Innovative Data Mining Techniques and Applications in Social Networks

Panteleimon Vikatos

A dissertation submitted
to the
Department of Computer Engineering and Informatics
University of Patras
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
Innovative Data Mining Techniques and Applications in Social Networks

Panteleimon Vikatos

A dissertation submitted to the
Department of Computer Engineering and Informatics
University of Patras
in partial fulfillment of the requirements
for the degree of
DOCTOR OF PHILOSOPHY

Supervisor
Associate Professor Christos Makris

Advisor
Professor John D. Garofalakis

Advisor
Professor Vasileios Megalooikonomou

Committee Member
Professor Sotiris Nikoletseas

Committee Member
Professor Ioannis Hatzilygeroudis

Committee Member
Assistant Professor Eleanna Kafeza

Committee Member
Associate Professor Kyriakos Sgarbas
I would like to dedicate this work to my future children.
Acknowledgements

Firstly, I would like to thank my supervisor Associate Prof. Dr Christos Makris from the Department of Computer Engineering and Informatics at University of Patras (UPatras) who gave me the directions to accomplish my thesis. This dissertation would not have been done without freedom, encouragement and full support of the last four years. I still remember the first discussion in his office:

- CM: "Pantelis, the topic of social computing is a new trend. I believe that we can introduce innovative data mining techniques to this direction."

Also, I would like to thank the members of the 3-member consultative committee Prof. Dr Vasileios Megalooikonomou and Prof. Dr John D. Garofalakis for the trust and their proposals and the other members of my 7-member examining committee, Prof. Dr Sotiris Nikoletseas, Prof. Dr Ioannis Hatzilygeroudis, Assistant Prof. Dr Eleanna Kafeza and Associate Prof. Dr Kyriakos Sgarbas, for the honor that I was given to participate in the examination of my doctoral dissertation.

I would like to thank Prof. Dr Fabrício Benevenuto from Universidade Federal de Minas Gerais (UFMG) who hosted me at LOCUS laboratory during my research mobility. It was my pleasure to work with him and his students, Johnnatan Messias and Manoel Miranda.

Many thanks to my colleagues from the Department of Computer Engineering and Informatics at University of Patras (UPatras) Prokopios Gryllos, Konstantinos Siaterlis and Alkiviadis Savvopoulos for the dedicated and professional collaboration.

More importantly, I thank my family and friends who never ceased to support me and to encourage me throughout this journey.
Abstract

Nowadays, people constantly use online social networking sites sharing content about their daily lives and things that happen around them. These systems have revolutionized the way we communicate, by organizing our offline social relationships in a digital form. Simultaneously, the social media serve the intentions of marketers to promote their brands and products due to this massive participation in these platforms and the endless potentials of improving their strategies for more effective marketing campaigns.

The selection of the targets, the diffusion of brand promotion messages and the place of advertising content in popular web pages constitute some of most significant objectives of marketers in order to gain the interest of potential customers.

Our research objective is to deal marketing campaign tasks using data mining techniques. We explore new methodologies in marketing campaign targeting, conversational and personalized advertising as well as the information propagation in Online Social Networks.

We introduce a methodology for calculating user influence and select targets of marketing campaigns using bridge participation in the evolving social graph. We analyze the improvement of social bots infiltration using automated communication skills and we introduce the conversational social bots as advertising content promoters that improve brand engagement.

We provide a novel methodology for personalized advertising using hotlink assignment. Our method enhances browsing experience and leads users to certain advertising content through hotlinks.

We also introduce a novel methodology to achieve information diffusion within a social graph that activates a realistic number of users. Our approach combines the predicted patterns of diffusion for each node with propagation heuristics in order to achieve an effective cover of the graph.
Our methodologies are useful to recommendation systems as well as to marketers who are interested to use data mining techniques to run effective marketing campaigns.

**Keywords:** Social Media Marketing, Link Prediction, Graph Mining, Sentiment Analysis, Social Bots, Personalized Advertising, Hotlink Assignment, Website Complexity, Information Diffusion
# Contents

1 Introduction ........................................... 1
   1.1 Motivation ........................................... 2
   1.2 Goals .............................................. 3
   1.3 Contributions ....................................... 4
   1.4 Dissertation Outline ............................... 5

2 Related Work ......................................... 7
   2.1 Online Social Network Influence and Diffusion ........... 7
      2.1.1 Influential Users ................................. 7
      2.1.2 Information Diffusion ............................... 8
      2.1.3 Automated Infiltration ............................. 10
   2.2 Link Prediction in Online Social Networks ............... 11
   2.3 Customized Advertising using Social Media ............... 12
      2.3.1 Browsing Enhancement ............................... 12
      2.3.2 Targeted Advertising ................................. 13
   2.4 Concluding Remarks .................................. 14

3 Marketing Campaign Targeting .......................... 17
   3.1 Introduction ......................................... 17
   3.2 Model Overview ...................................... 18
      3.2.1 Social Media Crawler ............................... 20
      3.2.2 Multilayer Social Graph Generation ............... 20
      3.2.3 Pairwise Feature Extraction ......................... 21
      3.2.4 Bridge Extraction .................................. 26
      3.2.5 Bot Detection ...................................... 27
      3.2.6 Campaign Targeting ................................ 28
   3.3 Experimental Results ................................ 29
      3.3.1 Implementation ..................................... 29
3.3.2 Results ......................................................... 30
3.3.3 Discussion ................................................... 33
3.4 Conclusions Remarks ........................................... 34

4 Bots Infiltration in Marketing Campaigns ....................... 37
4.1 Introduction .................................................. 37
4.2 Model Overview ............................................. 39
  4.2.1 Generation of Conversational Social Bots ................ 39
  4.2.2 Transform Bot to Conversational Bot .................... 40
  4.2.3 Crawling Targets ....................................... 41
4.3 Experimental Results ........................................ 44
  4.3.1 Implementation ......................................... 44
  4.3.2 Results .................................................. 45
4.4 Conclusion .................................................. 48

5 Personalized Advertising ........................................ 51
5.1 Introduction .................................................. 51
5.2 Model Overview ............................................. 52
  5.2.1 Generation of website’s graph ............................ 54
  5.2.2 Extraction of user’s preferences .......................... 54
  5.2.3 Creation of Hotlinks ..................................... 55
    5.2.3.1 Hotlink Assignment ................................. 55
    5.2.3.2 The Algorithm ....................................... 57
  5.2.4 Website reconstruction & Ads placing .................... 59
5.3 Experimental Results ........................................ 60
  5.3.1 Implementation ......................................... 61
  5.3.2 Evaluation Metrics ...................................... 61
  5.3.3 Results .................................................. 63
5.4 Concluding Remarks .......................................... 67

6 Pattern Based Diffusion of Marketing Campaign .................. 71
6.1 Introduction .................................................. 71
6.2 Methodology ................................................ 73
  6.2.1 Crawling Twitter Data ................................... 74
  6.2.2 User Profiling ......................................... 74
  6.2.3 Prediction of Retweet Pattern ........................... 75
  6.2.4 Pattern Based Diffusion ................................ 76
6.3 Estimating the Actual Propagation .......................... 78
6.4 Experimental Results ........................................ 80
   6.4.1 Implementation ........................................ 80
   6.4.2 Results .................................................. 81
   6.4.3 Discussion .............................................. 83
6.5 Concluding Remarks .......................................... 85

7 Concluding Discussion and Future Work 87

A Summary in Greek 91
   Α’.1 Εισαγωγή .................................................. 91
   Α’.2 Κίνητρο .................................................... 92
   Α’.3 Στόχοι ..................................................... 93
   Α’.4 Συνεισφορά ............................................... 94
   Α’.5 Στοχοποίηση χρηστών ................................. 95
   Α’.6 Διείσδυση Αυτοματοποιημένων Μηχανών ............... 98
   Α’.7 Εξατομικευμένη διαφήμιση ................................ 100
   Α’.8 Διάχυση πληροφορίας στα Κοινωνικά Μέσα ............... 104
   Α’.9 Μελλοντική Εργασία ................................... 107

Bibliography 109
# List of Figures

3.1 System Architecture: The independent tasks that our methodology contains starting with the crawling procedure and ending up to the campaign targeting. ........................................... 19  
3.2 Multilayer social network \( G = \{G_a; a \in \{T, F\}\} \), where T and F constitute Twitter and Foursquare, respectively. The layer with blue nodes represents the Twitter users and the layer with orange nodes the Foursquare ones. The dashed line connects the accounts in the different layers. ..................................................... 22  
3.3 Existence of inter multilayer path of length 2 between nodes \( u \) and \( v \). The node \( u \) has a following relation with node \( x \) in cluster \( C_{T_1} \) in Twitter. The nodes of cluster \( C_{T_1} \) have interlayer connections with the cluster \( C_{F_1} \) in Foursquare. The node \( w \) has friend relation with node \( v \) and thus \( IMP(u, v, 2) = 1 \). ....................................................... 27  
3.4 Combinations of Feature Sets ........................................ 32  
3.5 Scalability results in relation and mutual-follow graph ............... 33  
4.1 The Botometer scores of each category. The overall, content, sentiment and network value constitute the different dimensions of bot detection. 46  
4.2 Klout score for experimental active bots during 1 month. ................. 46  
4.3 Follow Ratio score for experimental active bots during 1 month. ........ 47  
4.4 Message Interactions for experimental active bots during 1 month. ....... 47  
4.5 Improvement % on Klout score between CSB and Baseline Bot ........... 48  
4.6 Improvement % on Follow ratio between CSB and Baseline Bot ........... 48  
4.7 Improvement % on Message Interactions between CSB and Baseline Bot 49  
5.1 Presentation of the system architecture and independent tasks that our methodology contains for website reconstruction and placement of advertising content. ............................................. 53  
5.2 Example of Hotlink Assignment ....................................... 58
5.3 HTTP archive record (HAR file) extracts the page load, the number of object request and the size in kilobytes of webpage. 

5.4 Improvement % on average path distance between $G$ and $G \cup PH$. 

5.5 Average path page load between $G$ and $G \cup PH$. 

5.6 Improvement % on average path page load between $G$ and $G \cup PH$. 

5.7 Average number of object requests in paths. 

5.8 Improvement % on average number of path requested objects between $G$ and $G \cup PH$. 

5.9 Average path page size between $G$ and $G \cup PH$. 

5.10 Improvement % on average path page size between $G$ and $G \cup PH$. 

5.11 Improvement % on average path page load in categories between $G$ and $G \cup PH$. 

5.12 Improvement % on average path page size in categories between $G$ and $G \cup PH$. 

5.13 Improvement % on average number of path requested objects in categories between $G$ and $G \cup PH$. 

6.1 System architecture of pattern based diffusion. 

6.2 Tweet Tree Diffusion Structure Patterns. 

6.3 The average rate of coverage. The x-axis presents the number of steps and y-axis the average rate of coverage based on 10000 rounds of simulation. 

6.4 Results of each heuristic simulation. The x-axis presents the number of steps and y-axis the average rate of coverage based on 10000 rounds of simulation. 

6.5 Quality of clustering. 

6.6 Independent cascade model. 

Α'.1 Αρχιτεκτονική συστήματος: Οι ανεξάρτητες λειτουργίες που περιλαμβάνει η μεθοδολογία μας ξεκινώντας από τη διαδικασία δημιουργίας του γραφήματος και καταλήγοντας στη στόχευση της καμπάνιας. 

Α'.2 Βελτίωση % Klout μεταξύ CSB και Baseline Bot. 

Α'.3 Βελτίωση % Follow ratio μεταξύ CSB και Baseline Bot. 

Α'.4 Βελτίωση % Message Interactions μεταξύ CSB και Baseline Bot. 

Α'.5 Παρουσίαση της αρχιτεκτονικής του συστήματος και των ανεξάρτητων εργασιών που περιλαμβάνει η μεθοδολογία μας για την ανακατασκευή ιστότοπου και την τοποθέτηση διαφημιστικού περιεχομένου.
Α’.6 Αρχιτεκτονική συστήματος διάχυσης βασιμένο σε πρότυπα . . . . . . 105
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Crawling Statistics</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Relation Graph (F1 score)</td>
<td>32</td>
</tr>
<tr>
<td>3.3</td>
<td>Mutual Follow Graph (F1 score)</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Social Bots Configuration</td>
<td>45</td>
</tr>
<tr>
<td>4.2</td>
<td>Descriptive statistics of social behaviour</td>
<td>45</td>
</tr>
<tr>
<td>6.1</td>
<td>F-Measure for each Classifier</td>
<td>76</td>
</tr>
<tr>
<td>6.2</td>
<td>Results of clustering $k = 4$</td>
<td>84</td>
</tr>
<tr>
<td>Α’.1</td>
<td>Διαμόρφωση μηχανών AI</td>
<td>99</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

"The man who moves a mountain begins by carrying away small stones."

— Confucius, The Analects

The popularity of Online Social Networks (OSN) is increasing each day as the current statistics show. More specifically, the third quarter of 2017, Facebook and Twitter have 2.07 billion\(^1\) and 330 million\(^2\) monthly active users respectively. People constantly use online social networking sites sharing content about their daily lives and things that happen around them. These systems have revolutionized the way we communicate, by organizing our offline social relationships in a digital form.

Simultaneously, the social media serve the intentions of marketers to promote their brands and products due to this massive participation in these platforms and the endless potentials of improving their strategies for more effective marketing campaigns. The selection of the targets, the diffusion of brand promotion messages and the place of advertising content in popular web pages constitute some of most significant objectives of marketers in order to gain the interest of potential customers.

This massive popularity of social media not only provides the opportunity to detect useful characteristics and patterns about users and their interconnections but also to implement algorithms and data mining techniques in order to provide a useful tool and efficient strategies in marketing campaigns.

The data mining and machine learning techniques in the field of social media marketing have gained the interest of marketers and researchers in order to develop


tools for calculating users influence, recognizing social bots, capturing the information diffusion, the extraction of users interest etc.

This dissertation focuses on this context. Our research objective is to deal with specific tasks of marketing campaigns using data mining approaches. We interest in the exploration of new methodologies in marketing campaign targeting, conversational and personalized advertising as well as the information propagation in Online Social Networks.

We hope our methodologies to be useful to recommendation systems as well as to marketers who are interested to use independently our proposed techniques or as a unified framework to run effective marketing campaigns.

In the next section, we present the main motivations of our work.

1.1 Motivation

Nowadays, the available data from Online Social Networks gives a unique opportunity to extract patterns related to users interests, their interconnections and the level of their influence to other users. We consider that due to budget and time limitation companies focus only to a specific set of audience, i.e, target group; and thus the information which is provided from social media is used to promote the advertising content to certain users pursuing the optimization of their campaign. Despite the power of the myriad data-collecting and data analysis tools, the procedure of targeting is limited to the selection of a set of social network users, i.e, influential; that have as current status high level of influence or their posts have high visibility and reproduction based on their previous history. The OSNs constitute an evolving phenomenon and thus the behaviour and impact of users are changing during the time. We consider that this information differentiates the decision making during a marketing campaign. Current efforts lack this dimension which motivates us to research metrics including the future status of the users. A manner to model a social network is the graph which the nodes and edges constitute the users and their relations respectively. We claim that machine learning in crawled social graph contributes to the development of prediction models about the structure of the network in the near future. We consider that the research objectives of this prediction can be translated to applications of data mining techniques such as classification and clustering. Furthermore, we hope the contribution of heterogeneous OSNs which extend the topological, behavioral and linguistic features to provide us with the ability to develop efficient AI tools for the prediction and advisory support for managing marketing campaigns.
1.2 Goals

Also, the available tools and methodologies of marketers lack of automated promotion and most of their efforts in social media marketing are manual. We support that the use of automated AI machines is appropriate to improve the massive handling of the content promotion and to be an immediate and independent procedure for conversational advertising which captures the users’ needs.

Furthermore, marketers tend to place advertisements (Ads) of their products in popular web pages, however, users are annoyed by the presence of advertisement banners during their browsing experience based on the fact that the most of them are irrelevant to their preferences. The available information from social media gives the ability to customize web pages content by extracting user interests in implicit or explicit manner. We consider that the enhancement of user satisfaction during browsing correlates to his/her intention to receive advertising content or to buy certain products. The lack of a holistic approach of website reconstruction that takes into consideration the enhancement of browsing experience, e.g, reduction of browsing steps or page load, the interest of the user and the place of personalized advertisements motivate us to a methodology for this purpose.

Another aspect of digital marketing attention is the process of the spreading of messages, i.e, campaign content in the social network. The information diffusion process that simulates the actual spread of word, has gained the research interest mainly using probabilistic models. A plethora of tools and models have been revealed trying to investigate the spread of information, e.g, virus contamination. In addition, there are probabilistic models that have been introduced to capture the maximization of influence spreading, addressing the problem of recognizing the $k$ initial users that will maximize propagation. However, the introduced models lack the handling of this procedure based on real patterns of diffusion. We consider a new approach in information diffusion with the effective use of a diverse set of data, i.e., the actual network structure, the behavioural metrics of the user, the text of the actual tweets for a specific topic for developing a realistic policy for a marketing campaign.

1.2 Goals

The main goal of this dissertation is the development of data mining techniques and creation of AI tools for the improvement of the decision making policies in marketing campaigns. We describe our objectives along 4 axes.

The first constitutes the suggestion of an alternative influence metric that takes into account the evolution of OSN and the change of users status.
The second is the analysis of social bots infiltration and the use of automated AI machines for marketing campaign policies.

The third is the proposal of a methodology of the automated placing of advertising content in a dynamic website using information from social media and providing enhancement of browsing experience.

The fourth axis contains the definition of heuristics that allows simulating the information diffusion in OSNs with a more realistic and easily defined manner using the historical propagation patterns of the users in Online Social Networks.

1.3 Contributions

This dissertation presents the following contributions regarding each of our previous goals.

The main points of our contribution can be summarized in the following sentences:

- We propose an alternative approach for calculating user influence, the bridge participation in the evolving graph. Considering the influence metric as well as the exclusion of users that constitute content polluters or social bots from the set of candidate targets, a weighted cost distribution strategy is introduced to run effective marketing campaigns.

- We analyze the improvement of social bots infiltration using automated communication skills and we introduce Conversational Social Bots as advertising content promoters that improve brand engagement.

- We introduce an efficient procedure of Ads’ placing in a website in terms of adaptation to the user needs and load performance in browsing paths. Our methodology adjusts a holistic procedure of website reconstruction using a generic scheme for personalization through social media and an algorithm of hotlink assignment.

- We develop a methodology that captures the spread of information that can possibly occur in a network in a realistic manner. We propose the Pattern Based Diffusion algorithm which explores the information propagation based on users’ history. A marketing campaign that gives incentives to the type of users we propose, can achieve similar results to a probabilistic model of diffusion.

The results presented in this dissertation are part of the following papers (in chronological order of publication):
Conferences:


Journals:


1.4 Dissertation Outline

This dissertation is organized as follows.

Chapter 2 briefly reviews related efforts in the literature along three axes. First, we discuss the methodology used by efforts that characterize social media users’ influence and infiltration. Second, we discuss the main efforts that have explored in
the evolution of social network through link prediction. Third, we refer studies that combine the enhancement of browsing experience with the personalized advertising.

In Chapter 3, we present our methodology for improving targeting of marketing campaigns using bridge prediction in communities based on the information of multilayer online social networks.

Next, Chapter 4 describes the usage of automated AI machines, i.e., social bots; as advertising content promoters and as a tool for brand loyalty enchantment.

In Chapter 5, we introduce a novel methodology for personalized advertising using hotlink assignment. We provide an automated procedure that places advertising content improving the success of the campaign.

Then, Chapter 6 introduces a novel methodology to achieve information diffusion within a social graph that activates a realistic number of users.

Finally, Chapter 8 concludes the dissertation and offers directions for future work.
Chapter 2

Related Work

"The aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error."
— Bertolt Brecht, Life of Galileo

In this section, we review the related literature along 3 axes. First, we discuss the methodology used by efforts that characterize social media users' influence and infiltration. Second, we discuss the main efforts that have explored in the evolution of social network through link prediction. Third, we refer studies that combine the enhancement of browsing experience with the personalized advertising. Finally, we present the concluding remarks.

2.1 Online Social Network Influence and Diffusion

The main focus of the decision making in marketing campaigns constitutes the selection of the targets that influence other users or spread on a large scale and range advertising content to the social network. Also, new efforts introduce the use of automated AI machines that improve the engagement of users with brands or be used to conversational advertising.

2.1.1 Influential Users

The study [28] is one of the first efforts for estimation of the social influence of users conducted on a large amount of data collected from Twitter; The authors suggest that high in-degree indicates low influence in terms of creating re-tweets and mentions. They also show that high influence is created through targeted efforts and is usually associated with a specific set of topics per influential user. The authors of
the study [119] try to identify the users who most influence others’ activity, based on data collected from social networking sites, using a non-standard form of Bayesian shrinkage. The results show that approximately one-fifth of a user’s friends influence their activity level on the site. More work on the identification of influential network nodes has been done in study [100] where the authors target enterprise brand-page communities by creating and analyzing an implicit network based on user interactions. In the work [73], a novel method for finding influential users using the link structure and the temporal order of information adoption is proposed. In a recent study [62], the detection of influential users is targeted as a discrete optimization problem and in effort [76] Leskovec et al. study information cascades; phenomena in which an action or idea becomes widely adopted due to influence by others; by developing a scalable algorithm and a set of techniques that illustrate the existence of cascades and measure their frequencies. Their experimental results suggest that cascade sizes are relatively small following approximately a heavy-tailed distribution. The frequency at which different cascade subgraphs occur depends on the product type; Furthermore, these frequencies reflect more subtle features of the domain in which the recommendations are operating. The user-to-user content transfer is the topic of the empirical study presented in effort [8], where the authors study the time-evolving social network of the massively multiplayer online world virtual world of game Second Life. Their results show that adoption rates quicken as the number of friends adopting increases and variable according to the connectivity of a particular user. In study [7], users who have been influential in the past and who have a large number of followers seem to generate the largest cascades.

### 2.1.2 Information Diffusion

In current scientific research, there is a huge interest in studying different types of information diffusion processes on large graphs and social networks. The most extensively reported influence diffusion models are Independent Cascade (IC) and Linear Threshold (LT). Based on the probabilistic theory and processes for information diffusion [106, 52, 54] Kempe et al. in [62, 63] it is proposed a greedy algorithm for influence maximization problem message spreading through a social network. In effort [62], the issue of choosing influential sets of nodes is described as a problem in discrete optimization. The optimal solution in this problem is NP-hard for most models that have been studied, including models motivated by applications to the marketing of Domingos and Richardson which initiate the issue of maximization the spread of influence as algorithmic problem [103, 39]. In study [103], the optimization
problem is solved based on a system of linear equations, while in the very general model of [39] the optimization problem cannot even be approximated to within a non-trivial factor. Furthermore, a plethora of studies have been proposed in influence maximization problem with more efficient heuristics [30, 31, 32, 64, 125, 69]. The main drawback of the most of the proposed greedy methods is the scalability of their greedy algorithms [30, 31, 32]. An alternative methodology by Kimura and Saito in [64] provides efficient algorithms for message spreading using shortest-path based influence cascade models. A plethora of extensions of the classic models of IC and LT are proposed in the field of the competitive influence [69, 13, 93, 16, 118]. A model of competitive influence which actually an extension of IC model is introduced from Bharathi et al. [13] providing a polynomial approximation algorithm for trees. Moreover studying models of dynamics Kostka et al. [69] introduce a competitive rumour spreading model comparing the computing complexity with IC and LT models. Using model dynamics and particular the voter model Pathak et al. [93] describe a generalization of the linear threshold model with multiple cascades while and Borodin et al. [16] extend the LT model in several different ways to model competitive influence diffusion. A different approach based on the epidemic model of SIS is introduced by Trpevski et al. [118] examining the dynamics in several types of graphs. Recently the focus of scientific research has gained the patterns and the structure of the information cascades processes on large graphs and social networks. Leskovec et al. [76] deal with information cascades developing a scalable algorithm to recognize the existence of cascades which are actually the adoption of the idea due to influence by others as well as to measure their frequencies and size which is small and follows a heavy-tailed distribution based on their study. In study [9], authors identify common patterns in tweets’ propagation and many of these patterns involved transmission with a third and fourth distance from the source user recognizing that the stronger ties are individually more influential. In effort [51] the authors study the cascades in social networks indicating that the spread of information could be modeled as an epidemic process. In addition, the result of their procedure shows that the vast majority of cascades are small and they could be described as simple tree structures. Authors in study [7], quantify influence on Twitter discovering that the largest cascades tend to be generated by users that are followed by a large number of users. Another work [80] combines the problem of link prediction in Microblogs with the notion of social distance based on the interaction patterns. The authors in [80] discover the preferable tendency of the agents to create ties with other agents who are close to them following the notion of homophily.
2.1.3 Automated Infiltration

The rise and strategies of automated accounts in social networks have gained the interest of researchers which focus on their detection, influence and analysis of their characteristics and behaviour for improving their infiltration in the social network. Create bots, artificial intelligence entities, able to pass by humans has long been a challenge for the AI community. Recently, social networks provide a perfect environment for bots to pass by humans. Many efforts showed that social networks, particularly Twitter, are full of socialbots [44, 43, 33]. The applications vary from commercial chatbots, i.e Pandora Bot \(^1\), to mass manipulation during political events [12, 68, 45]. Broadly speaking there are two existing kinds of efforts related to social bots: 1) identify and detect approaches and 2) deploy and measure approaches. The first attempt to massive detection of bots introduces a seven-month deployment of 60 honeypots discovering 36,000 candidate content polluters [75]. The provided strategies for social bot detection suggest machine learning techniques using behavioural and structural features [33, 11, 75]. In Darpa Twitter bot challenge [111] different teams focused on the identification of influence bots inside on the Twitter network. A recent effort [44] presented differences between content that human and social bot promote and provided a tool that incorporated their findings into a machine learning model. Another study [33] used supervised learning to classify between Twitter accounts in three categories - users, bots and cyborgs (users assisted by bots). A similar study [123] discovered that that influence metrics such as Klout \(^2\), number of followers and friends, constitute significant factors to predict user interaction with bots. Complementary to the detection of bots, the study [122] analyzed structure, behavior and linguistic characteristics of the social network to predict user’s susceptibility to bot attacks. Also, there are efforts on spam detection that end up getting some bots that attack hashtag spam, by trying to promote certain hashtags [11]. In addition, spammers improve their reputation in their network gaining followers using the process of reciprocal exchange of links between unrelated users (link farming) [50].

The drawback of this kind of approach is that it is, in general, biased towards one specific type of social bot, i.e., there might be advanced social bots in Twitter that are not easily identified in any of the described approaches and might be hard even for humans to tell if they are social bots or not. Aiming at filling this gap, the second set of efforts consists of deploying social bots and evaluate potential new functionalities to them to test what works and what does not work. The first attempt [88]

\(^1\)http://www.pandorabots.com
\(^2\)https://klout.com
presented a social bot which posted automated tweets, followed users on Twitter and described itself as a Brazilian journalist and thus achieved a high influence score. A current study [47] investigated and compared different strategies exposing Twitter’s vulnerability against large-scale social bot attacks that can affect both Twitter itself and services built on gathered from Twitter. A study [18] in Facebook OSN showed that bot accounts were able to infiltrate successfully the network on a scale which depends on users’ privacy settings.

2.2 Link Prediction in Online Social Networks

The structure of Online Social Networks changes due to new users are added and new relations are created. The evolution of the social network constitutes a factor that determines the policy of a marketing campaign due to the born of new targets and influential users that can improve the promotion performance.

The social network is modeled as a graph which nodes and edges are the user and relations between them respectively. The current interest in the social graph evolution is concentrated in the prediction of new links. Survey studies [78, 81] present the similarity-based algorithms, probabilistic models, and applications in link prediction. In addition study [79], provides the analysis of new perspectives, challenges and methods in link prediction containing unsupervised and supervised techniques. Many approaches have been proposed to handle link prediction based on probabilistic models [40, 60, 124], matrix factorization [87], kernel-based learning [77] and feature-based methods with unsupervised [78] and supervised techniques [2, 105, 131].

The first research approach in link prediction is presented in [78] in which different network topological metrics are extracted from the social network to be correlated with the generation of a link between social nodes. This unsupervised method shows that Adamic-Adar metric has the best performance in predicting links. An alternative approach based on supervised random walks is declared in [6] which combines the topology of the network with node and edge level attributes. A probabilistic graphical model for link prediction is proposed in [40] to recommend connections through heterogeneous social networks. Researchers in [2] present SVM as the most efficient classifier to predict links in co-authorship graphs with limited characteristics comparing supervised learning algorithms. The expansion of feature vector with sentiment features could improve the performance of link prediction in terms of F-measure as suggested in study [131]. Furthermore using the combination of location,
social and global features study [105] introduces a supervised method to predict links on location-based social networks.

Studies [112, 113, 57] present the concept of link prediction through another dimension using heterogeneous social networks. Study [57] proposes a set of features extracted from Twitter and Foursquare and uses the Random Forest classifier to solve the problem of link prediction. Also, study [58] combines node-based features such as reputation and optimism [107] with cluster-based meta paths features [114] extracted from multilayer networks.

2.3 Customized Advertising using Social Media

The placing of advertising content in web pages and platforms is a task that marketers care about. User satisfaction during browsing experience is related to the consumer' probability to buy a product and support a marketing campaign. We refer the efforts in browsing enhancement. Then, we discuss the main efforts that have explored in personalized advertising and studies related to personalization in information systems.

2.3.1 Browsing Enhancement

The topic of the improvement of user satisfaction during browsing has gained the interest of researchers. Algorithms and techniques contribute to the reconstruction of a site in order to reduce the time and steps that a user needs to reach his/her interest.

One well-known approach [96], i.e., hotlink assignment, models the website as a graph which node and edges are the webpage and the links respectively and proposes the placing of new links which connect popular web pages with its descendants reducing the distance from the home page. It presents the idea of a modification of the link structure of the website, minimizing the steps from homepage to popular pages using hotlinks. Using the concept of hotlink assignments a plethora studies propose [17, 36, 70, 49, 4] differentiation in the features and improvement in algorithms complexity.

There are 2 basic approaches for the hotlink assignment the clairvoyant user model and the greedy model. Using the clairvoyant model the study [17] discovers the upper and lower bounds complexities in order to assign the hotlinks in a website. The authors of this study model the website as a directed acyclic graph (DAG). However, considering the website as a tree, study [36] shows a polynomial complexity of the algorithm for assigning a hotlink which outperforms greedy approaches.

On the other hand, the running time of optimal solution for greedy models [49, 97, 98] is exponential in the depth of the tree and polynomial for trees of logarithmic
depth, e.g., binary trees. In addition, approximation algorithm has been proposed for the problem and an algorithm of 2-approximation in terms of the gain has been presented in [85]. However, the authors in study [41] provide a linear-time algorithm using an alternative approach with dynamic programming and operations such as node insertion, deletion, and weight reassignment. Regarding website complexity and page load time, studies [21, 22, 48] investigate the complexity of websites with browser-based active measurements and evaluate how page load times impact user satisfaction.

2.3.2 Targeted Advertising

Prior e-marketing research presents that personalized advertising content is more appealing as it relates to consumers preferences [53, 72, 120]. For example, a recent study demonstrates that advertisements which target user interest are more effective than general marketing strategies [120]. The rapid progress in the big data algorithms and dynamic targeting technology leads to the conclusion that personalization constitutes the future of online advertising and product promotion [38, 109]. Social media like Facebook extract the available information about demographics, personal interests, connections, past behaviors, and future activities in order to use it in complex targeting algorithms and provide personalized advertisements in the users' news feed [1]. The use of personalized advertising tools and techniques is increasing; however, there is evidence that consumers are prone to change their behavior realizing that the information system captures their preferences. For instance, in study [1] discovered sharp drops in click-through rates due to users' inconvenience realizing that personal data were tracked without their consent. Another effort [120] suggests that customized advertisements are efficient only when users feel confident about privacy issues. The main focus in order to specify user interest is the development of the personalization techniques. Personalization methodologies can be summarized in three main categories in the way that the necessary information is collected, i.e., explicit, implicit and hybrid. A plethora of web pages and information systems use explicit personalization and infer user interest based on predefined categories that the user should select. The main drawback of an explicit collection of user feedback is that it is not supported by all users. On the other hand, many studies use the implicit discovery of preferences which is fed by the browsing history of a user e.g., clicks or by the available information from social media. The implicit feedback of the user preferences based on the web usage and browsing, as study [86, 61] refers, contributes to the penalization in web search engines.
In addition, the extraction of data through social media such as post/tweets, hashtags, likes and users lists is beneficial to the re-ranking of the produced results of information retrieval models [26] as well as the expansion of queries in order to provide personalized content as study [132] proposes. Also, the combination of implicit and explicit methods leads to personalization scheme with high performance in capturing users needs as study [92] suggests.

2.4 Concluding Remarks

In this chapter, we provided related work regarding the contribution of data mining techniques and algorithms in the field of social media marketing.

The characterization of users and their ability to influence others or diffuse the necessary information on a large scale is significant for marketers in order to select the targets. Previous studies focus on the categorization of a user as influential only their current status in terms of topological features or historical data about adoption of their propagated information from other users. However, online social networks constitute a constantly evolving phenomenon and the user current status is changed. We hope our study to fill this gap developing prediction models for the evolution of the social graph and extracting alternative metrics for users influence evaluation. Furthermore, we concentrate on machine learning techniques and analysis of multilayer OSN extending and combining topological, behavioral and linguistic features in order to improve the prediction performance.

The most research efforts in the field of social bots are gathered in developing a model of identification of automated AI accounts and analyzing the vulnerabilities of online social networks from the massive generation of content polluters/spammers. Our study gives an alternative dimension of use social bots, firstly by analyzing the level of their infiltration based on advanced communication functionalities and proposing as a marketing tool for conversational advertising and brand engagement policy.

The current evolution of digital marketing was born new concepts about advertising. The available information from social media gives the ability to create customized advertising content for the users. There are many studies that deal with the extraction of implicit and explicit information from online social networks and create the profile of users interests. Also, algorithms and methodologies for enhancing browsing experience have gained current focus of researchers. However, there is a lack of studies that
combines the concept and proposing efficient place of personalized advertisements to information systems and web pages using information from social media.
Chapter 3
Marketing Campaign Targeting

"The secret of my influence has always been that it remained secret."
— Salvador Dalí

3.1 Introduction

Brands try to promote products and advertisements to potential customers that will be interested in their promotions and try to maximize marketing campaign performance using information about customer needs [35, 42]. Although companies spend billions of dollars on this direction, targeting individual customers with ever more accuracy, the direct marketing itself seems to be wasteful despite the power of the myriad data-collecting and data analysis tools at their disposal. A contemporary methodology is to spread messages, e.g., advertisements, promotions only to a few users in social networks which are characterized as influential [63, 65, 127]. Businesses try to identify influential users for propagating messages by looking in most cases, on users’ static profiles, consumer referral behavior and several antecedents; scarcity of the product, information value, and consumer’s need to reciprocate [67]. Also, they rely on metrics and web-services [101] which provide the current status of the influence of potential target. Furthermore, the influence of a user can be differentiated due to behavioral or topological change inside the online social network [28].

The prediction of this change provides the influence status of a user which is significant to decision making during a marketing campaign. On the topological dimension, each user creates or removes links with other users which in the case of Twitter means a new following relation or the end respectively. In fact, this problem;
the prediction of link formation; is generalized as the link prediction problem (predicting new edges in the social network graph) and constitutes a core problem in the domain of SNA by creating the fundamental task of modelling the way relationships in social networks evolve; which can be also characterized as the task of analyzing the proximity between nodes in a network. In modern social networks, link prediction models can be used to predict possible new connections constituting big platforms like Facebook and Twitter able to provide their users with meaningful recommendations [6, 40, 78, 105, 108, 131]. Additionally, there are few attempts that tackle link prediction in multiple networks [57, 74, 116, 58] considering extra link information from other OSNs.

In our approach, we combine topological, linguistic, behavioral and multilayer features in order to achieve better performance in the link prediction task. Our goal is to identify the users that constitute the seeds (influential users) for marketing campaigns by looking at participation rates in bridges (existed + predicted) between communities introducing pairwise features on the social network. Our methodology not only identifies potential influential users, which is beneficial for businesses propagating messages to promote brands and products but also demonstrates the impact of multilayer information for the prediction improvement. Also, the proposed strategy introduces the detection of automated AI machines, i.e., social bots; removing these accounts from the candidate targets. Our results are useful to marketers developing strategies in which they may adopt different associated costs for users that are characterized as influential and constitute intermediates for different communities.

The rest of the chapter is structured as follows. In Section 3.2, we provide an overview of the methodology that we propose which is separated in sub-modules. Section 3.3 overviews details of the implementation of the system and presents a reference to our experimental results and the discussion of our work. Finally, in Section 3.4, we declare the contributions of our effort.

3.2 Model Overview

We consider that in marketing campaigns the decision of the accounts/targets is crucial to the success of the promotion. We adopt the methodology of study [55] which proposes the bridge participation of communities as a metric for the influence of the nodes. Our model, which is presented in Figure 3.1, suggests the prediction of the bridge participation that connects different communities in the multilayer graph based on the fact that a node that is involved may be more influential than other
Figure 3.1: System Architecture: The independent tasks that our methodology contains starting with the crawling procedure and ending up to the campaign targeting, which actually isolated in the network. In terms of message propagation, the nodes that are participating at most in these bridges constitute the targets that can diffuse in an efficient manner the information, covering more parts of disconnected parts of the graph and enhancing marketing campaigns. In addition, our methodology excludes the accounts that constitute automated AI machines using a dedicated task of social bots recognition. Our main focus is to train and test an efficient classifier for link prediction in pairs of nodes between different communities.

A summary of the independent tasks that our methodology consists of is given below:

- **Social Media Crawling.** Initiating from a list of nodes (seed) from online social networks, we gather data about the ego-centric network for each user and discover new nodes in a Breadth-First search (BFS) approach of the multilayer graph. Also, we crawl the recent test/posts of each user.

- **Pairwise Feature Extraction.** Handling the texts of each user, we extract the sentiment, linguistic, popularity and inter-layers characteristics for each pair of the existing graph. Also, we the structural

- **Bridge Extraction.** Using a classification model for link prediction between the pairs of nodes, that we trained using the crawled data, we discover the evolution
of the graph in terms of the new links. The expansion of the graph with the new predicted links creates new communities and bridges are extracted.

- **Bots Detection.** The accounts which constitute automated AI machines are detected and removed from the selection process of the candidate targets.

- **Campaign Targeting.** The nodes are ranked by bridge participation. Nodes with high participation are characterized as influential and intermediates between communities, spreading messages in separated parts of the graph.

In the following subsection tasks and modules of our model are described in detail.

### 3.2.1 Social Media Crawler

The crawling procedure is handled by a dedicated crawler that allows sampling of the Twitter data in a manner that the necessary information for the following tasks is acquired. The crawling is separated in 2 axes. The first one is the gathering of the posts/tweets and the user characteristics that are introduced for the training of the prediction model. The second is the creation of the graph. The proposed crawler performs a Breadth-First search (BFS) starting from one (or multiple nodes) in order to generate the graph. We store the crawled data using a unique identifier for each user account. The seeds for the graph generation are derived from historical data of brands/companies about customers’ loyalty in specific products and needs. We store the crawled data using a unique identifier for each user account. Our collected dataset contains a large set of Twitter profiles including the lists of user ids that correspond to friends and followers along with a sample of each user’s public posts stream (tweets and re-tweets). Additionally, we detected the foursquare profiles of the user’s that publicly shared their location on Twitter via Foursquare.

### 3.2.2 Multilayer Social Graph Generation

A well-known way to model the interactions between different users in a social network is by using a graph $G(V, E)$ where the nodes in set $V$ represent users and the edges $E$ represent interaction. However, this representation focuses only on one kind of interaction. Multilayer networks have been studied for a wide range of practical applications including air travel and online multiplayer games [25, 115]. Despite the fact that modern OSNs are characterized by multiplexity [19] and the apparent need for high-resolution multichannel data when modeling large-scale networks [110] extracting knowledge from the structure of multilayer social networks for practical SNA
applications has not been a major focus so far. The topology of multilayer geo-social networks has been used for gaining insight about user interactions [57]. The authors in effort [66] provide a thorough discussion on the formalization of multilayer networks as well as on the topic of single to multilayer network generalization. Existing works have also focused on defining frameworks for the studying social links beyond individual social networks [57, 66].

A multilayer network is a pair of $M = (G, C)$, where $G = \{G_\alpha; \alpha \in \{1, 2, \ldots, m\}\}$ is a set of graphs $G_\alpha(X_\alpha, E_\alpha)$ and $C = \{E_{\alpha\beta}; \alpha \subseteq X_\alpha * X_\beta; \alpha, \beta \in \{1, 2, \ldots, m\}, \alpha \neq \beta\}$ is a set of interconnections between nodes of different layers $G_\alpha$ and $G_\beta$ with $\alpha \neq \beta$ and $m$ the number of layers. The interconnections $C$ are separated on intralayer and interlayer. The intralayer connections of layer $\alpha$ are the elements of set $E_\alpha$. The interlayer connections of layers $\alpha$ & $\beta$ are the the elements of the set $E_{\alpha\beta}$ ($\alpha \neq \beta$). We present a 2-layer graph in Figure 3.2, e.g., Twitter and Foursquare. The use of different layers not only increases the number of features but also reflects that the proximity on one social network may have an impact to a future linkage to another.

We consider two types of networks:

- The mutual-follow graph (MF graph) in which all edges are reciprocal.
- The relation graph (Rel graph) which is a generalization of the follow graph.

It is noted that we are treating the relation graph as undirected, thus an edge signifies only the existence of a relation (link) between two users (nodes). We widen the relation of each pair of users in other social media in order to obtain a multilayer social network. We use the names of extracted account in order to check their existence in the Foursquare \(^1\) social network. The common users/accounts in both social networks constitute the nodes of the multilayer social network. We note that this procedure is scalable to add more layers derived from other social media.

### 3.2.3 Pairwise Feature Extraction

We consider that the factors that are related to the possibility that a pair of user might have a link in the future are in four independently dimensions. One dimension is the topological proximity in the derived network from the social media. Related studies [2, 130] propose topological metrics and show that are powerful indicators of nodes proximity in a network. The second dimension constitutes the sentiment/linguistic metrics. Although, a pair of nodes has distance regarding their position the network,

\(^1https://foursquare.com/\)
3.2. Model Overview

Figure 3.2: Multilayer social network $G = \{G_a; a \in \{T, F\}\}$, where T and F constitute Twitter and Foursquare, respectively. The layer with blue nodes represents the Twitter users and the layer with orange nodes the Foursquare ones. The dashed line connects the accounts in the different layers.

the tweets in terms of topic or sentiment that this account posts can attract a user to follow. Another dimension is the level of influence that a user has already acquired. It is common sense that celebrities and well-known users are more likely to connect with other users based on their popularity. All above dimensions focus on the information that other social network can provide. Users participate in different social networks and the interaction in one can lead to the creation of a new link to another introducing significant inter multilayer metrics.

Topological Metrics

We provide the definitions of the topological metrics for a pair of nodes in graph $G$ as study [108] provides. We declare as $\Gamma(x)$ the set of neighbors of node $x$ and $|A|$ is the number of elements in set $A$.

$$\Gamma(x) = \{y| (x,y) \in E \lor (y,x) \in E\} \quad (3.1)$$

Common Neighbors (CN): A pair of users is more likely to have a link if they have many common neighbors. We define it as the size of the set of all common friends between $x$ and $y$ according to Equation 3.2.

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)| \quad (3.2)$$

Shortest Path Length (SPL): The distance between two users is related to the possibility that this pair will be connected. We extract all possible paths and we focus on the shortest path length is Equation 3.3.
3.2. Model Overview

\[ SPL(x, y) = d_G(x, y) \]  
(3.3)

Preferential Attachment (PA): The probability that a new link will connect \( x \) and \( y \) is proportional to \( |\Gamma(x)| \times |\Gamma(y)| \) which corresponds to the pairwise metric of Equation 3.4.

\[ PA(x, y) = |\Gamma(x)| \times |\Gamma(y)| \]  
(3.4)

Jaccard Coefficient (JAC): A metric that provides the the proportion of shared nodes between \( x \) and \( y \) relative to the total number of nodes connected to \( x \) or \( y \) as Equation 3.5 presents.

\[ JAC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \]  
(3.5)

Adamic Adar (AA): This measure refines the simple counting of common neighbors by weighting more heavily the less-connected neighbors defined in Equation 3.6.

\[ AA(x, y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\log |\Gamma(z)|} \]  
(3.6)

Similarity Pair Coefficient (SPC): The similarity of the pair coefficients refers to the estimate that derives from Equation 3.7.

\[ SPC(x, y) = 1 - |Coef[x] - Coef[y]| \]  
(3.7)

Ranking Pair Coefficient (RPC): The ranking of the pair coefficients refers to the summation of the coefficients of all the common neighbors of \( x \) and \( y \) as shown in Equation 3.8.

\[ RPC(x, y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} Coef_z \]  
(3.8)

Summation Pair Coefficient (SUPC): The summation of the pair \( xy \) simply refers to the summation of the coefficients of the nodes \( x \) and \( y \) as Equation 3.9 presents.

\[ \text{Rank}_{Coef} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} Coef_z \]  
(3.9)

The coefficient in the metrics SPC, RPC & SUPC can be one of the Centrality, Page-rank and Clustering based features as study [130] provides. The category centrality is further divided into the degree, betweenness and eigenvector centrality.
3.2. Model Overview

Sentiment & Linguistic Metrics

We introduce a second set of sentiment features proposed in [56, 108, 131] and extend it with a similarity of the linguistic analysis of users’ texts. The sentiment and linguistic metrics are topic independent contrary to similar studies [131, 104] that use topic affiliation to improve the performance of link prediction in social networks.

We define $P_i, N_i, O_i$ as the positive, negative and objective sentiment respectively for a set of hashtags of user $i$. The number of hashtags constitutes the union of the different sentiments, i.e., $H_i = P_i \cup N_i \cup O_i$. Also, we define $P_j, N_j, O_j$ as the set of users that expresses positive, negative objective sentiment in a a hashtag $h_j$ respectively.

Common Hashtags (CHA): We calculate the number of common hashtags among the nodes $x$ and $y$ as Equation 3.10 presents.

$$CHA(x, y) = |H_x \cap H_y|$$  \hspace{1cm} (3.10)

Occurrence (OCC): It is a boolean feature that describes the existence or not of common hashtags among $x$ and $y$.

Sentiment Adamic Adar (SAA): It describes the sum of the Adamic Adar distances for all common hashtags for which the users share the same opinion as Equation 3.11 shows.

$$SAA(x, y) = \sum_{h_j \in P_x \cap P_y} 1/\log |P_j| + \sum_{h_j \in N_x \cap N_y} 1/\log |N_j| + \sum_{h_j \in O_x \cap O_y} 1/\log |O_j|$$  \hspace{1cm} (3.11)

Sentiment Agreement (SAG): It calculates the number of common hashtags for which the two users expressed the same sentiment is presented in Equation 3.12.

$$SAG(x, y) = |P_x \cap P_y| + |N_x \cap N_y| + |O_x \cap O_y|$$  \hspace{1cm} (3.12)

Sentiment Disagreement (SDI): It provides the number of common hashtags for which the two users expressed opposite sentiment excluding hashtags for which the users expressed an objective opinion as Equation 3.13 refers.

$$SDI(x, y) = CHA(x, y) - SAG(x, y)$$  \hspace{1cm} (3.13)

Sentiment Aligned (SAL): It provides the ratio of sentiment agreement (excluding objective opinions) over the number of common hashtags between the two users as Equation 3.14 presents.
3.2. Model Overview

\[ SAL(x, y) = \frac{|P_x \cap P_y| + |N_x \cap N_y|}{|H_x \cap H_y|} \]  \hspace{1cm} (3.14)

Sentiment Misaligned (SMAL): It provides the ratio of sentiment disagreement over the amount of common hashtags as Equation 3.15 declares.

\[ SMAL(x, y) = \frac{|P_x \cap N_y| + |N_x \cap P_y|}{|H_x \cap H_y|} \]  \hspace{1cm} (3.15)

Sentiment Inverse (SIN): The sum of the inverses of the adoptions for each common hashtag for which the users share the same opinion as it is presented in Equation 3.16.

\[ SIN(x, y) = \sum_{h_j \in P_x \cap P_y} 1/|P_j| + \sum_{h_j \in N_x \cap N_y} 1/|N_j| + \sum_{h_j \in O_x \cap O_y} 1/|O_j| \]  \hspace{1cm} (3.16)

Sentiment Rarest (SRA): The adoption of the rarest common hashtag excluding hashtags for which the users expressed objective opinion as Equation 3.17 shows.

\[ SRA(x, y) = \min(\min_{h_j \in P_x \cap P_y} |P_j|, \min_{h_j \in N_x \cap N_y} |N_j|) \]  \hspace{1cm} (3.17)

Sentiment Mean (SME): The mean size of common hashtags by summing up the adoption of all hashtags for which the users share the same opinion and diving that by the sentiment agreement of the users as Equation 3.18 defines.

\[ SME(x, y) = \frac{1}{SAG} \times \left( \sum_{h_j \in P_x \cap P_y} |P_j| + \sum_{h_j \in N_x \cap N_y} |N_j| + \sum_{h_j \in O_x \cap O_y} |O_j| \right) \]  \hspace{1cm} (3.18)

Topic-SVO distance (ESVO): The Euclidean distance of the svo scores [56] of two users which is the sentiment volume objectivity value as Equation 3.19 presents.

\[ SVO(x, y) = \sqrt{\sum_{h_j \in H_x \cap H_y} (s_x(h_j) - s_y(h_j))^2} \]  \hspace{1cm} (3.19)

Linguistic similarity: We use the Linguistic Inquiry and Word Count (LIWC) [95] software which parses users’ text and assigns 80 features that include the linguistic and psychological use of language. The cosine distance of pair vectors formulates the metric as it is shown in Equation 3.20.

\[ Sim_{LIWC}(x, y) = \frac{\sum_{k=1}^{N} L_x[k] \cdot L_y[k]}{\sqrt{\sum_{k=1}^{N} L_x[k]^2} \cdot \sqrt{\sum_{k=1}^{N} L_y[k]^2}} \]  \hspace{1cm} (3.20)
3.2. Model Overview

**Popularity Metric**

The third set includes one metric, as it is presented in [55], which describes the similarity of influence for a pair of users based on Klout [101] which uses information from heterogeneous social networks and provides a value in range 1 to 100 as it is shown in Equation 3.21.

\[
Sim_{pop}(x,y) = \max(\text{klout}(x),\text{klout}(y)) \frac{|\text{klout}(x) - \text{klout}(y)|}{\text{Top_klout}}
\]  

(3.21)

The factor Top_klout constitutes the maximum Klout score of all users in the social graph.

**Inter Multilayer Metrics**

We use a meta path-based features as study [58] proposes. The traverse from one layer to another means the change of the relation. We consider that it is common a pair of nodes without a direct link or a path with a certain distance (hops) in one layer and thus we compute the distance between two users based on the clusters that the users belong in their interlayer connections. We compute for each pair all the paths from node \( u \) in \( G_\alpha \) to node \( v \) in \( G_\beta \) following the interlayer connections. Figure 3.3 describes an example of calculated a meta path. The set \( MP^{G_{\alpha\cup\beta}}_{u,v} \) includes the calculated distances of pair \( u,v \). We define 3 features which is actually boolean values of the existence of the meta-path of length \( k \) in set \( MP^{G_{\alpha\cup\beta}}_{u,v} \).

\[
IMP(x,y,k) = \begin{cases} 
1, & \text{if } k \in MP^{G_{\alpha\cup\beta}}_{u,v} \\
0, & \text{otherwise} 
\end{cases}
\]  

(3.22)

**3.2.4 Bridge Extraction**

We consider that the social graph evolves and new pairs of nodes connect and thus the nodes may acquire links and be intermediate between different communities during a time window. More specifically, the bridge extraction process contains 2 steps, i.e., the link prediction between different the pairs of the graph and the community detection.

We consider that the problem of prediction of a link in the future of a social graph belongs to a family of the prediction problems of binary classification. There are several feature-based classification methods [2, 78, 131, 105] that train and test supervised learning algorithms in order to create models for link prediction. Also, alternative approaches such as matrix factorization [87] and probabilistic models [40, 60, 124]
3.2. Model Overview

Figure 3.3: Existence of inter multilayer path of length 2 between nodes $u$ and $v$. The node $u$ has a following relation with node $x$ in cluster $C_{T_1}$ in Twitter. The nodes of cluster $C_{T_1}$ have interlayer connections with the cluster $C_{F_1}$ in Foursquare. The node $w$ has friend relation with node $v$ and thus $IMP(u, v, 2) = 1$.

have been presented as independent information systems or parts of a recommendation platforms.

The procedure is the same for the mutual following graph and the relation graph. Firstly, we calculate the topological, linguistic, sentiment, popularity and interlayer features for each pair. From each crawling, we obtained a progressively different snapshot of the same sub-network of the Twitter graphs. We used those snapshots to annotate the feature vectors giving a label $Y$ when a link exists in the future snapshot and a label $N$ otherwise.

As experimental procedure provides, the comparative study of well-known algorithms of supervised learning leads us to the selection of a classifier that achieves the best performance in terms of F-measure. In community detection dimension, there are several approaches as presented in [46] and [99] where the advantages and disadvantages of each method are documented. Methods and algorithms are discussed in [15, 91] for fast and efficient extraction of communities that use modularity as the measure of partition quality. The modularity-based criterion quantifies the quality of the divided communities in which dense internal connections indicate, in terms of the Twitter network, groups of users with strong ties regarding follow relationships.

3.2.5 Bot Detection

Current studies [44, 121] declare the rise of social bots in the Twitter environment. This task focuses on the detection of social bots which constitute automated AI machines that interact through tweets, links and available functions provided by social
3.2. Model Overview

The recognition of social bots is important for the performance of the marketing policy due to bot accounts even if can easily transform to influential users [88, 47] in many cases are dedicated as content polluters [11] and thus the selection of these special cases as our marketing targets can lead to unexpected outcomes. In Twitter-sphere, there are a plethora of strategies related to bot recognition using behavioral and structural features [33, 11, 75]. We use the well-known web service, Botometer [37], to automated recognition of the bots.

3.2.6 Campaign Targeting

The goal of marketing campaigns is to diffuse its message (e.g. promotions, advertisements) to a very broad audience. Marketers rely on demographics in order to select the target group, for instance, the gender, location and age. The demographic status of the target group is derived from the historical data about the preferences and needs of each customer. However, the budget limitations lead to a different choice about the subset of the audience that the marketing campaigns are promoted. Hence, marketing campaigns focus on influential users [63, 65, 101, 127]. In our methodology, we expand this concept in two axes. Firstly, we use as additional information the detection of bridges in the communities. Also, we exclude social bots from the set of candidate targets. The campaign strategy involves the identification of nodes with high brand loyalty by using fine-grained attributes derived from historical data of previous campaigns. These nodes are the initial seed for the crawling procedure, therefore, the graph that is created, which constitutes a sub-graph of the social network, consists of dedicated and possible customers. The nodes of the graph are ranked by bridge participation. The bridges constitute predicted links between nodes of different communities. Regarding budget, each campaign sets a cost for sending advertisements and promotions to certain targets, and once the budget is spent the campaign no longer exists. The cost is distributed in accordance with the ranking of each user. We define the set of users $U$ in the graph $G$ as $U^G = S_{Inf}^G \cup S_{others}^G$. The set $S_{Inf}^G$ includes the k-top influential while set $S_{others}^G = U^G - S_{Inf}^G - S_{bots}^G$ includes the remaining nodes from the selected communities removing the set of bots. In addition, the budget for each set is formed using a coefficient $w$ that declares the ratio of budget dedicated to each set. More specifically the daily budget for each set is $B_{Inf}^G = \frac{w B}{D}$ and $B_{others}^G = \frac{(1-w)B}{D}$ where $B$ is the budget, $D$ is the duration in days and $w = \{0, 1\}$. The distribution of the budget is different in the two sets as presented in Equation 3.23.
3.3. Experimental Results

Cost$(u_i) = \begin{cases} 
\frac{B_{g|i}}{N_{g|i}} x_i^{e-\lambda} e^{-\frac{\lambda}{1-D_{\text{other}}}}, & \text{if } u_i \in S_{Inf}^G \\
\lambda_i e^{-\lambda} B_{g|\text{other}}, & \text{otherwise}
\end{cases} \quad (3.23)

In $S_{Inf}^G$ the daily cost is uniformly divided among users in the duration of the campaign. Whereas the cost in the remaining users follows a Poisson distribution with parameter $\lambda > 0$ that denotes the average number of users per influence interval and $i$ is equal to the ranking of each user based on bridge participation.

3.3 Experimental Results

3.3.1 Implementation

We contracted our experiments using data that we collected from the public Twitter API using the tweepy\(^2\) tool. We used pyfoursquare\(^3\) which is a Python wrapper for the data request from the Foursquare social network. We implemented the multilayer graph using the procedure described in Algorithm 1. For each user of Twitter, we searched his/her existence in Foursquare. Our approach provides us an intersection of two social networks through a subset of users who are active on both and have chosen to share their check-ins on Twitter. Furthermore, we recognized the accounts which constitute automated AI machines through an available API, i.e., Botometer [37]. The Botometer provides a JSON file of the each requested account with content, sentiment, network, and overall score in range $0 – 100\%$. We consider that a Twitter account with a score over than $60\%$ has an unusual behavior and constitutes a social bot.

We used networkx\(^4\) to model the Twitter and Foursquare graph and extract necessary centrality metrics. We used the Louvain Modularity algorithm [15] from the community. We handled the sentiment analysis of the tweets with the module textblob\(^5\). Also, we implemented the machine learning algorithms using scikit-learn\(^6\) to train all our classifiers except for the XGBoost classifier that was trained using the xgboost\(^7\) package.

Firstly, we crawled the Twitter network at two separate timeframes. Each crawling lasted 4 days. From each crawling, we obtained a progressively different snapshot of

\(^2\)http://www.tweepy.org
\(^3\)https://pypi.python.org/pypi/pyfoursquare
\(^4\)https://networkx.github.io
\(^5\)https://textblob.readthedocs.io
\(^6\)http://scikit-learn.org
\(^7\)https://pypi.python.org/pypi/xgboost
3.3. Experimental Results

<table>
<thead>
<tr>
<th>Property</th>
<th>Rel graph</th>
<th>MF graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>V_{GT,F}</td>
<td>$</td>
</tr>
<tr>
<td>$</td>
<td>V_{GT\setminus F}</td>
<td>$</td>
</tr>
<tr>
<td>$</td>
<td>E_{GT,F}</td>
<td>$</td>
</tr>
<tr>
<td>$</td>
<td>E_{GT\setminus F}</td>
<td>$</td>
</tr>
<tr>
<td>Avg tweet count per user</td>
<td>565</td>
<td></td>
</tr>
<tr>
<td>Avg Twitter friends count per user</td>
<td>3090</td>
<td></td>
</tr>
<tr>
<td>Avg hashtag count per user</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Avg klout score per user</td>
<td>37.4</td>
<td></td>
</tr>
</tbody>
</table>

the same sub-network of the multilayer graph. We explored 34,180 and 4,861,413 nodes and edges respectively. After one month period, we crawled again the same graph. Hence, the second snapshot contained the same number of nodes, however, 78,516 new edges were added. The data extraction of the Foursquare social network constitutes the next step. More specifically, we searched the crawled tweets of each user to find if they shared a Foursquare check-in. Our crawling procedure calculated a rate of 23% of users that are active on both social networks. We created a link between the node in the layer of Foursquare using the list of friends for each node.

3.3.2 Results

To provide a better insight into the quality of our crawled data, we provide in Table 3.1 a set of network properties corresponding to the relation and mutual-follow graph that was generated. $|V_{GT,F}|$ denotes the set of nodes on Twitter only while $|V_{GT\setminus F}|$ the set of nodes, which exist on both social networks, i.e., Twitter and Foursquare. We define $|E_{GT,F}|$ and $|E_{GT\setminus F}|$ the set of edges on Twitter only and on both OSNs respectively. In addition, more general crawling statistics are presented like the average Klout score of our users and the average number of tweets and hashtags that we collected for each user.

Furthermore, the Figure 3.4 describes the feature sets. We examine the performance of our classifiers on two different datasets, the first produced using the relation graph (Rel graph) and the second using the mutual-follow graph (MF graph). We used a balanced dataset including the same number of instances for each label of the relation and mutual follow graph. Each instance of the dataset includes the pairwise metrics of a pair of the graph and a label that denotes the existence or not of a link in the second snapshot.


3.3. Experimental Results

Algorithm 1 Multilayer Graph Generation

1: **input** id of influential user $\#id_{Inf}$, threshold T
2: **output** The sample relation graph $G^\text{Rel}_{T \cap F}$, The sample mutual-follow graph $G^\text{MF}_{T \cap F}$
3: Found_nodes = {$\#id_{Inf}$}
4: Explored_nodes = {}
5: while length(Explored_nodes) < T do
6:   for each $u_k \in$ Found_nodes do
7:      identify set of Twitter friends of user $u_k$, Friends$^T_{u_k} = \{f_{T1}, f_{T2}, \ldots, f_{Tj}\}$
8:      identify set of Foursquare friends of user $u_k$, Friends$^F_{u_k} = \{f_{F1}, f_{F2}, \ldots, f_{Fj}\}$
9:      for each $f_j \in$ Friends$^T_{u_k}$ do
10:         if shared_check_in($f_j$) == shared_check_in($u_k$) == True then
11:            Found_nodes = Found_nodes $\cup$ $f_j$
12:            $E_T = E_T \cup (u_k, f_j)$
13:            if $f_j \in$ Friends$^F_{u_k}$ then
14:               $E_F = E_F \cup (u_k, f_j)$
15:         end if
16:      end if
17:   end for
18:   Explored_nodes = Explored_nodes $\cup$ $u_k$
19: end for
20: end while
21: $G^\text{Rel}_{T \cap F} = G^\text{MF}_{T \cap F} = {}$
22: for each $e_{i,j} \in E_T$ do
23:     if $i \in$ Explored_nodes && $j \in$ Explored_nodes then
24:        $G^\text{Rel}_T = G^\text{Rel}_T \cup e_{i,j}$
25:     if $e_{i,j} \in E_F$ then
26:        $G^\text{MF}_F = G^\text{MF}_F \cup e_{i,j}$
27:     end if
28: end if
29: end for
30: for each $m_{e_{i,j}} \in G^\text{Rel}_{T \cap F}$ do
31:     if $m_{e_{i,j}} \in G^\text{Rel}_{T \cup F}$ then
32:        $G^\text{MF}_{T \cap F} = G^\text{MF}_{T \cap F} \cup m_{e_{i,j}}$
33:     end if
34: end for
3.3. Experimental Results

We compared several classifiers that have been proposed in previous studies [6, 40, 78] for supervised learning in link prediction using topological and additional sets of features [105, 108, 131]. Also, we used the XGBoost classifier for gradient boosting, as study [29] presents, which in our knowledge has never been used before for link prediction in multilayer social networks. We trained and tested using K-Fold Cross-Validation (K=10). Table 3.2 and Table 3.3 exhibit the performance of each classifier using alternative sets of features in relation and mutual-follow graph respectively. We infer that the F1 score of the classifiers varies in a range of 0.765 - 0.86 in the Rel graph and 0.702 - 0.907 in the MF graph. More specifically using only the topological features, we show that Random Forest outperforms the other classifiers when evaluated on the relation graph whereas XGBoost outperforms the other classifiers when evaluated on the mutual-follow graph. The addition of features related to linguistic, sentiment, and popularity creates a slight improvement in the performance. We observe while increasing the feature set with the complete set evaluated the XGBoost acquires the highest score. While most classifiers accuracy changes slightly with the addition of more classification features Naive Bayes seems to not benefit by the addition of features other than the structural ones. Furthermore, we presented the scaling ability of each classifier in the Figures 3.5(a) and Figure 3.5(b) using the topological features of Set1. The success rate of our classifiers converges in the area near the 70% – 90% of the training examples.
3.3. Experimental Results

(a) Scalability in Rel graph  
(b) Scalability in MF graph

Figure 3.5: Scalability results in relation and mutual-follow graph

Table 3.3: Mutual Follow Graph (F1 score)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
<th>Set 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.904</td>
<td>0.905</td>
<td>0.906</td>
<td>0.906</td>
<td>0.906</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.799</td>
<td>0.798</td>
<td>0.798</td>
<td>0.805</td>
<td>0.805</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.907</td>
<td>0.908</td>
<td>0.908</td>
<td>0.909</td>
<td>0.91</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.827</td>
<td>0.826</td>
<td>0.826</td>
<td>0.831</td>
<td>0.833</td>
</tr>
<tr>
<td>SVM</td>
<td>0.853</td>
<td>0.864</td>
<td>0.858</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.702</td>
<td>0.702</td>
<td>0.702</td>
<td>0.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>

3.3.3 Discussion

According to our experimental procedure, the XGBoost acquires the highest score for the prediction of new links in the graphs. We denote the accurate prediction of the specific day/time that a link will have existed depends on the sampling procedure of the temporal network and the training of the classifier. In our case, the time window constitutes the 1 month period. We used the trained model of XGBoost classifier to explore the link predictions between the pairs of nodes in 2 communities with the largest number of nodes since it was the one achieving the best score, to predict the creation of potential new edges between the two communities with most nodes. The 2 most-crowded communities contain 7706 and 6111 nodes respectively.

We focus only on the predictions that constitute bridges of these communities and the nodes that participate in them. We choose the MF graph in order to exploit stronger ties and decompose the parts of the graph where high connectivity occurs. Our choice is also supported by our experimental results which suggest that higher
prediction performance can be achieved when exploiting the mutual following topology. As the last step, we run the community detection algorithm on the resulting graph, the original mutual-follow graph extended with the predicted bridges. We apply Louvain’s Modularity algorithm [15] for community detection. The selected communities of our experiments are expanded 37% and 34.9% respectively in terms of the number of nodes. Also, we use the bot detection procedure to evaluate the examined nodes. Our results show that 427 nodes of the top-sized community and 462 nodes of the second community have Botometers’ overall score over than 0.6. These accounts are removed from the following steps of the analysis. The bridge detection rate in the case of the 2 most-crowded communities is approximately 0.8%.

We target users according to their participation in bridges that already exist in the graph or are created during the marketing campaign. As study [55] presents in order to check the correlation between bridge participation and impact in social media, we use the well-known metric Klout that has been used in many studies [3, 101] that measure user’s influence in a social network. We correlate the bridge participation with the Klout score for each user using the Spearman’s correlation coefficient. We scale the correlation analysis increasing the number of users in the ranking list. In the case of top 500 users bridge participation indicates significant positive linear correlation with the Klout metric (the correlation score is close to 0.9), while even in the case of 1000 top users the positive correlation remains strong (the correlation score is close to 0.6). In both cases, the p-value is approximately zero which indicates high statistical confidence in those results.

3.4 Conclusions Remarks

In this chapter, we present an innovative methodology to improve the efficiency of marketing campaigns through the identification of nodes with high bridge participation in the evolving social network. We focused on the detailed description of the methodology for predicting the future linkage of users in the social network, which constitutes the basic component of our work.

The main points of our contribution can be summarized in the following sentences:

- Our strategy suggests a weighted cost distribution based on the level of influence and introduces the removal of users that constitute content polluters or bots from the set of candidate targets.
• We use an alternative approach for calculating user influence, the bridge participation in the evolving graph, to run effective marketing campaigns.

• We perform a comparison study of well-known machine learning classifiers for link prediction in multilayer social networks.

• We combine topological, behavioral, linguistic and interlayer features of multilayer Online Social Networks in order to predict the future links with the highest performance.

Our novel methodology ultimately aims to provide a good basis for studies that want to establish a possibly iterative procedure through which they can continually train and improve on the results and generalization ability of an existing predictor.
Chapter 4

Bots Infiltration in Marketing Campaigns

"I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

— Alan Turing, Computing machinery and intelligence

4.1 Introduction

Online social networks (OSNs) constitute an immediate platform where people post and interact each other via exchanging message, sharing content and follow other users. The massive popularity of online social media provides the opportunity to researcher and developers not only to detect useful characteristics and patterns about users and their interconnections but also to implement advanced data mining, natural language processing, artificial intelligence and machine learning techniques for automation and human-computer interaction purposes.

Brands/companies promote their products and advertising content in the wide audience and thus the development of automated AI machines (bots) contributes to achieving this objective by capturing users’ interests and propagating advertising content in a massive way. The applications of bots vary from commercial chatbots, i.e, Pandora Bot ¹, to mass manipulation during political events [12, 68, 45]. Particularly,

¹http://www.pandorabots.com
the Twitter OSN is becoming a suitable place for the proliferation of bots, i.e., social bots; and many studies [33, 11, 75] focus on the automated detection of these accounts. Also, the use of advanced AI techniques and the combination with efficient strategies can transform a social bot to an influential user [47, 88]. The first attempt [88] presented a social bot which posted automated tweets, followed users on Twitter and described itself as a Brazilian journalist and thus achieved a high influence score. A current study [47] investigated and compared different strategies exposing Twitter’s vulnerability against large-scale social bot attacks that can affect both Twitter itself and services built on gathered from Twitter. A study [18] in Facebook OSN showed that bot accounts were able to infiltrate successfully the network on a scale which depends on users’ privacy settings. Furthermore, Facebook introduced new functionality of their Facebook Messenger application-bots \(^2\). FB messenger bots constitute an interactive assistant that provides functionality to its users, from shopping assistance, news reading to getting information about weather through the personalized weather forecast.

In this chapter, we investigate if social bots that chat, i.e., establish automatic conversations through direct messages, can help social bots to infiltrate into social networks. In previous studies, the functionality of chat was not enabled and did not test assuming that this could count negatively to the bot and even be used by users to decide to follow back or not. We consider that the interactive actions support our goal to infiltrate to social network and create an engagement between bot and users. To do that we deployed a set of social bots on Twitter that use state of the art techniques to chat and we compared its infiltration with existing social bots. Our study introduces social bots that are able to start a conversation, send a comment and reply to direct messages. Our methodology consists of creating social bots accounts with advanced AI conversational functionality investigating the influence in terms of Klout, message interactions and the number of followers. Also, we propose a dedicate crawler that gathers available information of user and extracts the users profile and interests.

The most research works in the area of social bots are gathered in the creation of prediction models of these kinds of accounts and in the discovery of the vulnerabilities of online social networks from the massive generation of content polluters/spammers. Our contribution gives an alternative dimension of handling social bots. Firstly, we investigate the level of infiltration of social bots with high-level communication functionalities, i.e., Conversational Social Bots (CSB). Secondly, we propose this kind

\(^2\)https://messenger.fb.com/
of bots as a marketing tool for conversational advertising and brand engagement policy. We consider that the analysis of social bots infiltration and the use of AI machines can introduce a new direction in social media marketing.

The rest of the chapter is structured as follows. In Section 4.2, we provide an overview of the configurations and functions of the baseline social bots (BSB). In addition, a set of additional functions contributes to the transformation of the baseline social bots to conversational using the ability of immediate response to direct message and replies to tweets and comments. Section 4.3 provides an overview of the implementation of bots and presents a reference to our experimental results. Finally, in Section 4.4, we discuss the contributions of our effort.

4.2 Model Overview

Our methodology introduces dedicated AI machines (social bots) as content promoters. We explore the level of infiltration of social bots and their communication performance in terms of the influence, the ratio of followers/following and the message interactions. Our methodology begins with the creation of a set of social bots. We focus only on Twitter due to its API that gives us the flexibility to implement social bots with all available actions. The configurations and differentiation in Twitter’s actions are plenty considering the number of options. According to the study of Freitas et al. [47], which provides $2^k$ factorial experiments [20], the configuration and activity of social bots determine the level of their infiltration in Twitter. However, their analysis lacks advanced communication skills which are necessary in order to promote advertising content through conversations.

4.2.1 Generation of Conversational Social Bots

We inspired by study [47] which describes in detail infiltration strategies of social bots and presents a comparison study between an alternative set of attributes in order to examine the factors of the most influential bots. Each bot account defines a set of characteristics in profile settings, and a set of functions in activity level. In terms of gender, a female bot tends to be more influential. Also, a high activity strategy that implements an action, e.g., post a tweet, follow an account; between 1 and 60 minutes reflects the level of visibility and increases the opportunity of other users to interact with the bot. In addition, the bots that post automated tweets are more reliable than the bots with only re-posting function. Regarding synthetically tweet generation, a Markov chain [10, 59] constitutes a mathematical model used to create
text that looks similar to the text contained in a sample set of documents. Another factor which affects how social bots are able to engage socially is the set of target users with whom the social bot attempts to interact. The study [47] infers that the bots that target to users who are interested in a common topic (hashtag) it is possible to connect each other. Based on the above infiltration strategies we define social bots with the following characteristics which constitute our baseline.

**Profile settings:** We declare the configurations of the baseline bots.

- name: English female first and last name. The email account is the merging of the first and last name using Gmail domain.
- profile picture: A female image with all copyrights.
- biography: A small description of job, interests and location (e.g, Journalist - Covering World News around the world, London, England).
- background: A cover image related to the interests that biography describes.

**Baseline Activities:** A randomized process which selects one of the following actions in a time window of 1-60 min.

- Posting Tweets: 1) Re-posting recent tweets from a set of pre-selected hashtags. 2) Generating automated text with a Markov chain.
- Follow users in a group: The bot expands its network by following users in a certain target group.

### 4.2.2 Transform Bot to Conversational Bot

Twitter OSN provides functions for users to like, retweet or comment a specific post. Also, a new function is the direct messages in other user. We consider that by using these extra functions to social bots improve their communication skills by proving to followers that bots are real and active. We differentiate from the the baseline bots using additional functions as the following lists presents. We call the proposed social bot as Conversational Social Bot (CSB).

**Additional Activities :**

- Likes : The bot selects a person from the corresponding group and performs a 'Like' on a recent post.
- Retweets : The bot post a retweet from a user in the corresponding group.
4.2. Model Overview

- **Post Trend Tweet**: The bot selects randomly a trending tweet from a specific location.

- **Comments**: The bot sends a comment to a recent post from a random user of the corresponding group.

- **Direct messages**: The bot replies to direct messages from other users. Also, bot can initiate a conversation though direct message to followers.

The activities of the CSB are distinguished in 3 main parts. The first part constitutes the handling of the response of direct messages. The second part is the activation of certain daily actions. In our case, the social bot selects randomly one of its followers and initiates a conversation, igniting a potential controversy regarding his/her own tweets. Also, the social bot selects randomly a trending tweet from a specific location. The third part includes the service of Twitter-based actions, i.e., like, post, follow, comment and retweet.

The creation and automation of CSB are provided in Algorithm 4.1. The algorithm starts with the creation of 2 threads for the service of daily actions and the response of direct messages respectively. At lines 2 and 3, the algorithm performs the initialization step of threads creation that are running independently. At line 5, an action from our list is selected. From line 17 to 35, the algorithm acts relative to the selected baseline activity and the additional one.

### 4.2.3 Crawling Targets

This task is responsible for targets discovery. Initially, each social bot has a list of users ids which are derived from historical data of brands/companies about customers’ loyalty in specific products and needs. The proposed crawler focuses on 2 axes. First, it determines the follow/unfollow operation of the social bot and performs a Breadth-First search (BFS) starting from one or multiple users. The list of users follower ids is stored and the recent tweets are gathered for each one. Second, it extracts the user interests form the tweets and their hashtags. The procedure of interest extraction includes a preprocessing module, i.e., stop-words removal, tokenization and stemming; and a Topic modeling algorithm which in our case is the Latent Dirichlet Allocation (LDA) [71]. The crawler compares the extracted topics with product-oriented list of keywords. Users that have at least one match with the product keywords receive a follow request from the social bot.
Algorithm 4.1 Conversational Bot Generation

1: **input:** id\_Bot, group, actions\_List
2: create\_Thread(DM\_Listener,id\_Bot)
3: create\_Thread(Daily\_Actions,id\_Bot)
4: **while** true **do**
5: \hspace{1em} s = select\_Action(actions\_List)
6: **switch** s **do**
7: \hspace{2em} **case** "Like"
8: \hspace{3em} User\_id = rand\_user(group)
9: \hspace{3em} recent\_post = select\_post(User\_id)
10: \hspace{3em} do\_Like(id\_Bot,User\_id,recent\_post)
11: \hspace{2em} **case** "Repost"
12: \hspace{3em} TH = select\_hashtag(Trending)
13: \hspace{3em} Post = select\_post(TH)
14: \hspace{3em} do\_Repost(id\_Bot,Post)
15: \hspace{2em} **case** "Post\_Markov"
16: \hspace{3em} Post = Markov\_msg()
17: \hspace{3em} do\_Post(id\_Bot,Post)
18: \hspace{2em} **case** "Follow"
19: \hspace{3em} User\_id = rand\_user(group)
20: \hspace{3em} **if** User\_id \notin Follow\_List(id\_Bot) **then**
21: \hspace{3em} \hspace{3em} Follow\_List = User\_id \cup Follow\_List
22: \hspace{3em} **end if**
23: \hspace{2em} **case** "Comment"
24: \hspace{3em} User\_id = rand\_user(group)
25: \hspace{3em} recent\_post = select\_post(User\_id)
26: \hspace{3em} msg = Chatbot(recent\_post)
27: \hspace{3em} do\_Comment(msg,id\_Bot,User\_id,recent\_post)
28: \hspace{2em} **case** "Retweet"
29: \hspace{3em} User\_id = rand\_user(group)
30: \hspace{3em} Post = select\_post(User\_id)
31: \hspace{3em} do\_Retweet(id\_Bot,Post)
32: \hspace{1em} wait\_time = rand(1,60)
33: \hspace{1em} sleep(wait\_time)
34: **end while**
Procedure 4.1 DM Listener
1: input id_Bot
2: while true do
3:   Message, Sender = Incoming()
4:   if Message is not empty then
5:     Response = Chatbot(Message)
6:     Send_DM(Response, id_Bot, Sender)
7:     Message = ∅
8:   end if
9: end while

Procedure 4.2 Daily Actions
1: input id_Bot
2: while true do
3:   s = select_daily_Action()
4:   switch s do
5:     case "Conversation"
6:       User_id = rand_user(group)
7:       msg = Markov_msg()
8:       Send_DM(msg, id_Bot, User_id)
9:     case "Trending Post"
10:        TH = select_hashtag(Trending)
11:        random_Tweet = select_post(TH)
12:        do_Post(random_Tweet)
13:        wait_time = rand(1200, 1440)
14:        sleep(wait_time)
15:   end while
4.3 Experimental Results

We conducted an experimental procedure in order to evaluate our methodology in terms of the level of influence, infiltration and message interaction of Conversational Social Bots. We focus only on Twitter due to its API that gives us the flexibility to implement social bots with all available actions.

4.3.1 Implementation

We created a set of accounts on Twitter. The social bots were differentiated in their functions as Table 4.1. We contracted our experiments using Twitter API and tweepy\(^3\) which is python module for activation of the automated functions. The social bots are topic oriented selecting 2 different topics, i.e., sports, news. Also, we defined the bots to follow a number of initial accounts and to have a number of posts and likes. More specifically the social bots followed each other. Our experiment started with 10 re-posting trend tweets and 10 like for each bot. We implemented the automated generation of posts using the Markov chain\(^4\) that was created based on 20 articles from CNN news and Telegraph related to the topic of interest. The chat function implemented using PyAIML\(^5\) which is an interpreter package for AIML, the Artificial Intelligence Markup Language with a standard set of conversations (standard). Every time the bot received a message the listener of Twitter streaming API activated the PyAIML module. The response of PyAIML was sent back to the initial sender, maintaining a proper conversation. In terms of comments, the bot selected a recent post from a random user of the corresponding group and isolated the post’s text field. Then the module of PyAIML produced the comment based on the extracted text. Furthermore, bots initiated a conversation by selecting randomly followers and igniting a controversy regarding on their own tweets. The bots were alive over a period of 30 days. In addition, all social bots "sleep" between 00:00 - 02:00 and starts again 10:00-12:00 GMT+2 time zone, simulating the expected downtime of human users. We checked to what extent social bots can gain popularity and influence in the Twitter social network. We use the following metrics to quantify how successful a social bot is.

- Follow ratio: This is a standard metric for estimating the ratio followers and friends of a user.

\(^3\)http://www.tweepy.org  
\(^4\)https://pypi.python.org/pypi/markovify  
\(^5\)https://pypi.python.org/pypi/PyAIML
Table 4.1: Social Bots Configuration

<table>
<thead>
<tr>
<th>Bot Type</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSB</td>
<td>Like Post Markov Follow Comment Retweet Chatbot</td>
</tr>
<tr>
<td>Baseline + L/R/C</td>
<td>Like Post Markov Follow Comment Retweet</td>
</tr>
<tr>
<td>Baseline + L</td>
<td>Like Post Markov Follow</td>
</tr>
<tr>
<td>Baseline</td>
<td>Post Markov Follow</td>
</tr>
</tbody>
</table>

Table 4.2: Descriptive statistics of social behaviour

<table>
<thead>
<tr>
<th>Bot Type</th>
<th>Likes</th>
<th>Tweets</th>
<th>Mentions</th>
<th>DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSB</td>
<td>195</td>
<td>700</td>
<td>60</td>
<td>42</td>
</tr>
<tr>
<td>Baseline + L/R/C</td>
<td>200</td>
<td>658</td>
<td>59</td>
<td>17</td>
</tr>
<tr>
<td>Baseline + L</td>
<td>218</td>
<td>614</td>
<td>57</td>
<td>19</td>
</tr>
<tr>
<td>Baseline</td>
<td>20</td>
<td>605</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

- Klout score: Klout score is a popular measure of the influence in Twitter which ranges from 1 to 100, with higher scores implying a higher online social influence of a user.

- Number of message-based interactions: We measure the number of times other users interact with a social bot through messages (tweets), @mentions the bot, replies to the bot, retweets or favorites a tweet posted by the bot.

We monitored their interactions with other users over a period of 30 days.

4.3.2 Results

Firstly, we examined the level of bots’ recognition from normal account/users regarding their replies in direct messages. We use Botometer [37] which constitutes a well-known online web service for bot detection [44, 37, 121]. The Botometer API provides a json with content, sentiment, network and overall score in range 0-100 for a Twitter account. Based on the overall score the created bots are not recognizable. Figure 4.1 shows that the CSB bot performs the lowest overall score. We infer that bots’ topology in the network reveals their automation due to the fact that bots overcome 56% and the baseline bot reaches the 70%. Content and sentiment attributes seem to be the factors in which the baseline bot reveals with scores 66% & 75% respectively. Also, we monitored the activity during our experimental period and Table 4.2 provides a description about the average likes, tweets, mentions and direct message for each type of bot.
4.3. Experimental Results

Figure 4.1: The Botometer scores of each category. The overall, content, sentiment and network value constitute the different dimensions of bot detection.

Figure 4.2 presents the fluctuation of Klout of each bot during the 1 month period. We can consider that there is no significant difference between bots until the end of the fifth day. However, during the 10-30 day the Conversational Social Bot reaches significantly higher score which equals 46.4.

Follow ratio expresses the level of visibility for each bot. We infer from Figure 4.3 that follow ratio value of CSB is higher than others after 2 weeks activity. The CSB has the highest follow ratio score that equals 0.65. The lowest follow ratio is derived from the baseline bot with 0.29.

The metric of message interactions is related to the communication of bots with other users. Figure 4.4 provides the comparison between the different type of social
4.3. Experimental Results

Figure 4.3: Follow Ratio score for experimental active bots during 1 month. We infer that the existence of direct message functionality improves the communication skills of the social bot and thus the performance of this metric is much higher than other. The CSB reaches the 125 total interactions at the end of 1 month period.

Figure 4.4: Message Interactions for experimental active bots during 1 month. We infer that CSB has better performance in terms of Klout, follow ratio and Message Interactions. In Figures 4.5 - 4.7 are presented the percentage of improvement in comparison to study of Freitas et al [47]. We consider that there is a significant improvement. The 10 day constitutes the peak of Klout improvement with 37%. Also, the major difference between CSB and baseline is at the end of 30 days reaching 123%
of better performance. The message interactions for CSB is about four times more than the baseline.

Figure 4.5: Improvement % on Klout score between CSB and Baseline Bot

Figure 4.6: Improvement % on Follow ratio between CSB and Baseline Bot

4.4 Conclusion

In this chapter, a collection of Twitter social bots was generated, whose final followers were attached to their network at their own discretion. The group of the Twitter bots, performed a large amount of human replicated activities in their network, in an attempt to socialize through tweets, comments and likes and even direct messages.
4.4. Conclusion

It is believed that the certain amount of actions of the social bots are insignificant in regards to the vastness of Twitter’s network. Moreover, the tweets and statuses of the social bots are either re-tweets of existing public content of other users or products of a word-generating model which operated on a tweet based corpus. In the concern of content credibility, user mentions and URLs were excluded from the tweet posting process, to establish spam free and user-friendly posts. In addition, connecting to the local network of the bot as a user constitutes a multifaceted decision. Users evaluated the subject matter of the bots in comparison to their interests and performed a follow or an unfollow action accordingly. The experimentation concluded after the preordained timespan of thirty days, ergo the deactivation of the Twitter accounts. The usernames of the bots and the connected users will remain undisclosed.

We investigate a new functionality for social bots in Twitter. Our research question was to check if the ability to chat would not affect negatively the performance of a social bot, but it turns out that it affected positively. Existing behavior-based detection strategies consider that social bots might not interact too much. We focus on the analysis of the influence in terms of Klout, message interactions and number of followers. We show that "off-the-shelf" libraries to allow a social bot to automatically chat can be deployed and be used to favor social bots and make it even better to infiltrate twitter.
Chapter 5

Personalized Advertising

"Many a small thing has been made large by the right kind of advertising."
— Mark Twain

5.1 Introduction

In the digital marketing, it is difficult to cover the needs of all users when the number of pages and categories in a dynamic website increase. Many websites contain hundreds or thousands of different categories, creating difficulties to find the page that the user wants, degrading the quality of the website. Social media constitute an alternative way to discover users’ preferences and can be used as a communication tool for the customers [27]. This capacity is basic when companies merge social media into their marketing strategies to improve customer engagement [126]. Simultaneously, the online engagement in a website content and browsing experiences are related to the positive or negative reactions to advertisements [23] and thus the reconstruction of a website constitutes a significant factor to enhance browsing combined by the accession of advertisement content that users need. The goal of reconstruction is not only to reach popular pages in fewer steps starting from the homepage improving the browsing experience but also to cover users’ needs about new products and maximizing marketing campaigns. A well-cited methodology is the use of additional links, i.e., hotlinks [41, 17, 97, 98], that connect popular webpages with its descendants, reducing the distance from the home page.

This chapter deals with personalized advertising inside of a dynamic website. We consider that user satisfaction during browsing experience is related to the consumer’ probability to buy a product and support a marketing campaign. Each advertising
content (Ads) constitutes an item, i.e., banners, videos, links related to products and marketing campaigns. Our methodology focuses on this content creating links (hotlinks) in the following dimensions. One dimension is to enhance the browsing experience of users to reach in fewer steps their necessary information which actually connected with advertising content. Also, the proposed methodology captures user preferences through information from social media. Furthermore, our methodology places the advertising content with respect to website complexity examining factors as page loads in the decision.

The main points of our contribution can be summarized in the following sentences:

- We introduce an efficient procedure of Ads’ placing in a website in terms of adaption to the user needs and load performance in browsing paths.

- Our methodology adjusts a holistic procedure of website reconstruction using a generic scheme for personalization through social media and an algorithm of hotlink assignment.

The rest of the chapter is structured as follows. In Section 5.2, we provide an overview of the methodology describing each independent task. We present the algorithm of hotlink assignment in Section 5.2.3.2. Section 5.3 provides an overview of the implementation of the system for modules and sub-modules respectively and references our experimental results. Finally, in Section 5.4, we discuss the strengths and limitations of our approach and we conclude with an outlook to future contributions.

5.2 Model Overview

Our methodology introduces an efficient placing of advertising content, i.e., banners, videos, links in a dynamic website. The procedure includes a holistic website reconstruction though hotlink assignment as studies [83, 84] provide. We present a hotlink assignment algorithm in order to enhance the browsing experience based on the reduction of the distance between popular pages and the target page. We differentiate this procedure linking pages with context similarity and supporting the hotlinks between pages that reduce the overall complexity of the web browsing in terms of load and number of object requests. Also, we exploit the available information from social media in order to personalize the creation of hotlinks based on user interests.

A summary of independent tasks that our methodology consists of is given below:
5.2. Model Overview

Figure 5.1: Presentation of the system architecture and independent tasks that our methodology contains for website reconstruction and placement of advertising content.

- **Generation of website’s graph.** We create a directed graph \( G(V, E) \) to model the website where nodes and edges are pages and links to the website respectively. We declare a popularity attribute in each node and 2 independent weights, i.e., a context-similarity and a page load gain weight. Also, based on the semantic information the category of each webpage is extracted, creating a list of categories for the whole website.

- **Extraction of user’s preferences.** This task is dedicated to producing a list of categories derived from user’s texts in social media. We use the topic modelling approach gathering the users’ posts on social media. This task is operated only when the user is active on the social network.

- **Creation of Hotlinks.** A randomized algorithm generates additional links for a node to one of its descendants to reduce the distance reduction from the home page to the page target improving the browsing experience.

- **Website Reconstruction & Ads Placing.** This task is responsible for website reconstruction using a ranking metric on the generated graph with hotlinks taking into consideration the popularity of pages, the context-similarity, the load performance and the personal interests.

In the following subsection tasks and modules of our model are described in detail and Figure 5.1 presents the system architecture.
5.2.1 Generation of website’s graph

We model the website as a graph $G(V,E)$. Our procedure uses crawling which is handled by a dedicated crawler that is developed for the task and that allows sampling pages and links in a manner that network properties are preserved and can be used in our modeling procedure. The crawling is oriented to discovering new nodes in a Breadth-First search (BFS) approach. We introduce an attribute calculating the popularity of each page/node and taking into account a page’s clicks as well as ranking in the social media trend list. Also, we calculate the complexity of each webpage in terms of page load. The advertising repository (Ads repo) constitutes a list of dynamic content, i.e., banners, videos, links etc. related to products and marketing campaigns. Each element of the list is introduced in the graph as a unique node. The pre-existing advertisements remain their links in the graph. The new ones will acquire the same links as the Ads with relevant context similarity. We use a term based text similarity approach, i.e., TF-IDF (Term Frequency-Inverse Document Frequency) using cosine similarity as Equation 5.1 presents; to compare the semantic context of the webpages on the site.

$$ sim(d_i,d_j) = \frac{\sum_{k=1}^{n} d_{i,k}d_{j,k}}{\sqrt{\sum_{k=1}^{n} d_{i,k}^2} \sqrt{\sum_{k=1}^{n} d_{j,k}^2}} \quad (5.1) $$

where $d_{i,k}$ is the TF-IDF weight of term $k$ in document $i$.

Simultaneously to the generation of graph our crawling procedure extracts the semantic information of the webpages to map each page to a category. We gather the list of categories which is the necessary information in our personalization scheme as it is described below.

5.2.2 Extraction of user’s preferences

In this section, our personalization approach is described. Our goal is to extract the preferences of a user based on the provided data of social media and especially Twitter. One approach is to use a Topic modeling algorithm such as Latent Dirichlet Allocation (LDA) [71] introduced the raw text of the posts/tweets. The LDA is a topic modeling procedure which discovers underlying topics in a set of documents and infers word probabilities in topics. An alternative approach is presented in the paper [84] proposing a personalization scheme which has been inspired from study [82] in which a graph of categories is used depicting the current user’s preferences in search engine results in a query. Also, another approach to infer the interests in the Twitter.
5.2. Model Overview

A social network is presented in [14] that discovers the preferences using a list-based methodology of the subscribed users in this organizational feature.

We adopt the method of [84] using a category graph for each user. Each node in the graph is a category and the weighted link between nodes reflects the correlation of these categories to the users’ texts.

The existence of the node is the graph is related to the Wu&Palmer metric (Equation 5.2) which calculates the semantic similarity using the depths of two synsets in the WordNet taxonomies [129, 94], along with the depth of the LCS (Least Common Subsumer) between the category of the webpage and LDA produced topic

\[
wup(s_1, s_2) = \frac{2 \cdot \text{depth}(\text{LCS})}{\text{depth}(s_1) + \text{depth}(s_2)}
\]

where \(s_1\) and \(s_2\) are the synsets for a pair of terms. The weight in link is calculated by the Equation 5.3,

\[
w(e(u, v)) = \sum_k \sum_l (\text{LDA}_r(t_k)[\text{LDA}_r(t_l) - \text{LDA}_r(t_k)])^{-1}
\]

where \(t_k\) is the topic \(k\) that exceeds threshold \(T\) and belongs to the category \(u\). \(\text{LDA}_r(t_k)\) and \(\text{LDA}_r(t_l)\) constitute the ranking of topic \(k\) and \(l\) by using LDA (Latent Dirichlet Allocation) procedure respectively. The weights in links are updated when a new sample of texts is mined for a specific user. A preprocessing procedure initiates the collection of the users’ texts including stop-words removal, tokenization and stemming. Each word in the extracted topics is compared to categories and the category with the maximum similarity is stored if it exceeds a threshold \(T\).

5.2.3 Creation of Hotlinks

5.2.3.1 Hotlink Assignment

The hotlink assignment has been a topic of interest for many studies [96, 36, 17, 97]. The reconstruction of a website is created using additional links concerning the popularity of the webpages. The scope of hotlink assignment algorithms is to optimize an objective function related to the distance between the homepage and the popular webpages through browsing. At most of the studies the website is modelled as a tree \(T = (V, E)\) in which \(V\) is the set of webpages and \(E\) is the set of links. Each leaf-webpage contains a weight corresponding to the popularity of the webpage. However, the studies [4, 84] propose an alternative methodology that represents the

\(\text{synsets} : \text{set of cognitive synonyms each expressing a distinct concept}\)
whole website as a directed acyclic graph (DAG) using the criteria of popularity and context-similarity of websites in the decision of adding extra links from non-popular pages. Study [84] proposes the combination of webpage traffic through clicks with the context similarity of trend topics in social media. The introduction of information of trend topics is necessary due to the fact that the webpages, with high traffic in the past, may have obsolete and outdated context. Under these circumstances, the popularity of each node is formed as Equation 5.4 presents.

\[
\text{pop}_i = \begin{cases} 
\text{rankFact} \times \text{simFact} \times \text{clicks}_i, & \text{if } \text{wup}(tt_j, C_i) > T \\
\text{clicks}_i, & \text{otherwise} 
\end{cases} 
\]  
\begin{equation}
(5.4)
\end{equation}

where \( \text{simFact} = (1 + \text{wup}(tt_j, C_i)) \) and \( \text{wup}(tt_j, C_i) \) is the maximum Wu & Palmer similarity [129, 94] between each trend \((tt_j)\). The \( \text{rankFact} \) is the Twitters’ trend ranking and the \( \text{clicks}_i \) is the i’s webpage’s clicks. The \( T \) constitutes a predefined similarity threshold of the website categories and the trend topics of Twitter. In the case that a webpage has different context with all topics of Twitters’ ranking list the popularity of the page is related only to number of clicks.

The algorithm of hotlink assignment discovers all paths between a random page (source) of the graph and a page from the POP set (target) calculating the popularity of the paths. We differentiate from the previous studies [4, 84] using the additional criterion of webpage complexity in terms of page load of the browsing path. The Equation 5.5 describes the path metric

\[
\text{Path}_H A_j = w_1 \sum_{i \in \{ \text{path nodes} \}} \frac{\text{pop}_i}{\# \text{in} \_ \text{edges}_i} + w_2 \sum_{i \in \{ \text{path nodes} \}} \frac{\text{node} \_ \text{PL}_i}{\# \text{out} \_ \text{edges}_i} 
\begin{equation}
(5.5)
\end{equation}

where \( \# \text{in} \_ \text{edges}_i \) and \( \# \text{out} \_ \text{edges}_i \) declares the number of incoming and outgoing links of the node \( i \) respectively. The factors \( (w_1, w_2) \in [0, 1] \) declare the weight of each criterion in the calculation. The popularity and the page load of each node is defined as \( \text{pop}_i \) and \( \text{node} \_ \text{PL}_i \) respectively.

The nodes of the graph are separated in 2 sets the Pop and NonPOP nodes. The candidate hotlinks are assigned between NonPOP nodes, that exist in the path with the maximum \( \text{Path}_H A_j \), and the target node. Contrary to the study [84] we introduce 3 criteria to define the final selection of hotlinks. Firstly, we examine if the distance between the target and source is reduced. Then, we check the semantic similarity between target and NonPOP page. Finally, we use the criterion of the page load in case of common similarity. The Example 5.1 presents a detailed depiction of our hotlink assignment approach.
5.2. Model Overview

Example 5.1: Hotlink assignment

As Figure 5.2 presents the path \((A,D)\) is a path between a source and a target page, where \(A\) (blue color) is the source page, which is randomly selected, and \(D\) (green color) is the target page, which is a page from the POP set. In this example, there are two distinct paths that start from node \(A\) and end at node \(D\), the \((A,B,C,D)\) and \((A,E,F,G,D)\). The algorithm calculates the popularity of the two paths according to Equation 5.5. The number inside of the node provides the sum of normalized popularity and page load for each page.

We calculate the popularity of each path

\[
\text{Path}_{HA}^{A\rightarrow B\rightarrow C\rightarrow D} = 12 + 6 + 2 + 4 = 24
\]

\[
\text{Path}_{HA}^{A\rightarrow E\rightarrow F\rightarrow G\rightarrow D} = 12 + 3 + 1 + 1 + 4 = 21
\]

The \((A,B,C,D)\) path is the most popular path, thus this path is selected. Node \(B\) belongs to \(NonPOP\) set. The nodes that have the same nearest common ancestor (NCA) with \(B\) and be in \(NonPOP\) set constitute the candidate nodes for hotlink assignment. In this case, node \(H\) and \(B\) are the candidate nodes. Let node \(B\) be more semantically similar to node \(D\) than \(H\). Connecting node \(B\) with node \(D\) will reduce the path. Thus, a hotlink is created (red arrow at Figure 5.2) from node \(B\) to node \(D\). If nodes \(B\) and \(H\) had the same context similarity, then the path page load would have been taken into consideration.

5.2.3.2 The Algorithm

We adopt the approach of study [84] to create the hotlinks. However, we differentiate in the selection of the candidate paths using the metric of \(Path_{HA}\) and the additional criterion of path load for the final placing of the hotlink. The Algorithm 5.1 provides a detailed description. The input of the hotlink assignment algorithm constitutes the weighted graph as it is generated by the crawling procedure of the website. Also, the nodes of the graph are marked based on the users’ preferences. More specifically, an additional binary attribute defines the existence of webpage context in the user’s interest. Also, we denote a set of procedures which are introduced in the algorithm definition. The procedure \(\text{Rand}(X)\) returns a random selection of an element of the set \(X\) using the uniform distribution. The \(\text{BFS}(K)\) uses the Breadth-first search to traverse a graph, finds all paths from a starting node \(K\) and returns a list of all paths. The procedure \(\text{max\_Score}\) calculates the metric \(Path_{HA}\) for all paths between the source and target and returns a list with the maximum score paths. The procedure \(\text{minPath}(G,v,u)\) estimates the reduction of distance between nodes \(v\) and \(u\) of graph \(G\). The procedure \(\text{maxCSim}(S,v)\) calculates the semantic
5.2. Model Overview

(a) Node A (blue color) is the source page and node D (green color) is the target page.

(b) Candidate path (red) for hotlink assignment is \((A, B, C, D)\).

(c) Candidate path (red) for hotlink assignment is \((A, E, F, G, D)\).

(d) Hotlink is assigned between node \(B\) and \(D\).

Figure 5.2: Example of Hotlink Assignment

similarity between each node of set \(S\) and the node \(v\) and returns a list with the highest similarity nodes. The procedure \(\text{lessLoad}(S, v)\) compares the page load of the between each node of set \(S\) and the node \(v\) and returns the node with the lowest
5.2. Model Overview

page load.

Algorithm 5.1 Hotlinks Assignment

1: input $G(V,E)$ initial graph
2: output $H(V,E)$ graph with hotlinks
3: $H = G$
4: while $NonPOP \neq \{\}$ do
5:     source = Rand($V$)
6:     target = Rand($POP$)
7:     Paths = BFS(source)
8:     for each $p \in$ Paths do
9:         if $p :: source -> target$ then
10:             $Path_{s->t} = Path_{s->t} \cup p$
11:     end if
12: end for
13: $CandPath = \text{max}_\text{Score}(Path_{s->t})$
14: for each node $\in$ $CandPath$ do
15:     $G_{temp} = G \cup (target, node)$
16:     if $\text{minPath}(G_{temp}, source, target) < \text{minPath}(G, source, target)$ then
17:         $candNodes = candNodes \cup node$
18:     end if
19: end for
20: $y = \text{maxCSim}(candNodes, target)$
21: if $|y| > 1$ then
22:     hotNode = lessLoad($y, target$)
23: else
24:     hotNode = $y$
25: end if
26: $H = H \cup (target, hotNode)$
27: end while

5.2.4 Website reconstruction & Ads placing

A slightly adjusted PageRank (Aux) is used to rank the nodes of the generated graph with hotlinks. The differentiation with the previous studies [4, 84] constitutes the combination of the incoming links with similar context and lower page load nodes. We formulate the recursive formula for each webpage $i$ as Equation 5.6 presents to rank the nodes of the graph.

$$Aux(i) = Aux(i) + \frac{\sum_{v \in In(i)} Simetric(v) \ast [sim(v,i) + PL_{gain}(v,i)]}{\#\text{outgoing links of } v}$$ (5.6)
where $In(i)$ is the incoming webpages to page $i$, $Simetric(i) = \frac{q}{\#\text{webpages}} + (1 - q) \times Aux(i)$ is a weight of page $i$ that represents the involvement of the page in contrast to the whole website, where $q$ is a dumping factor, $sim(v, i)$ is the context-similarity of webpages $v$ and $i$ and $PL_{gain}(v, i)$ is the page load gain of webpages $v$ and $i$. Equation 5.7 presents the gain moving from page $v$ to page $i$, i.e., we track the page load difference and present it as positive or negative gain.

$$PL_{gain}(v, i) = \begin{cases} \text{PageLoad}(i) - \text{PageLoad}(v) / \text{Max}_{\text{PageLoad}}, & \text{PageLoad}(i) \geq \text{PageLoad}(v) \\ 0, & \text{else} \end{cases}$$

Initially, the $Simetric$ is equal to $\frac{1}{\#\text{webpages}}$ and $Aux$ is equal to 0. The output of the slightly adjusted PageRank is a list of links in a descending order which is used to reconstruct the website and placing the advertising elements. We declare that this procedure does not remove links or nodes from the website. The reconstruction is beneficial in the following cases:

- New advertisements are introduced in the repository.
- Webpages and Ads are becoming popular and trend topics.
- User changes his/her preferences.

The personalization scheme provides a list of categories for each user that maps his/her interests. Each node includes as a binary attribute that it is named $Pers$, i.e. $Y/N$, which defines that the category of the node is relative to the user preferences based on threshold $T$. We check each hotlink from the descending list which is produced from the slightly adjusted PageRank ($Aux$) and we add only the hotlinks that connect node with context relevance to the users’ needs.

We use the crawling procedure (BFS) in the website checking for each traversed node the list of the hotlinks and the attribute $Pers$. The new link is added only if $Pers = Y$. The extraction of new hotlinks depends on the change of user preferences and the frequency of the update of the repository of the advertisement content.

### 5.3 Experimental Results

We conducted an experimental procedure in order to evaluate our methodology in terms of browsing steps and path complexity. We compared the crawled graph with the updated one with hotlink assignments.
5.3.1 Implementation

Our system was implemented in Python 2.7. We contracted our experiments using data that we collected from the public Twitter API using the tweepy\(^2\) tool, and we performed topic modeling on the tweets using LDA\(^3\). We used NLTK\(^4\) module for preprocessing and haralyzer\(^5\) module for extraction of HAR files. To implement the various network related feature extraction methods we used networkx\(^6\).

Our research objective was to examine the methodology to real cases. Hence, we used websites related to news that have a plethora of different categories and many links inside. We aggregated the selected information of 2 news websites, i.e., CNN, BBC news. We preferred only the English version of the sites for sake of simplicity. However, our methodology can be easily adapted to multilanguage environments. We used dedicated crawlers with a Breadth-First search (BFS) in parallel and gathered the context, links and external Ads from each webpage. Based on the fact that the access to websites’ server log files is not allowed for our purpose we used the web service Alexa\(^7\) in order to extract the popularity in terms of clicks for each webpage. Also, we used the HTTP archive record (HAR file) as Figure 5.3 presents to extract the page load, the number of object request and the size in kilobytes of each webpage.

Furthermore, we simulated the users entering the website using a number of predefined Twitter ids. Firstly, we checked if these accounts correspond to automated AI machines (bots) or not using the Botometer [37] which constitutes a well-known online web service for bot detection [44, 37, 121]. The Botometer API provides a json with content, sentiment, network and overall score in range 0-100 for a Twitter account. We removed the case that an overall score over that 60% due to artificial and unusual behavior. Also, we checked if the account for the selected Twitter id is active. We gathered the recent 3200 tweets for each active user. In addition, we performed topic modeling on the tweets using LDA\(^8\) using the default options.

5.3.2 Evaluation Metrics

In this section, we introduce a set of evaluation metrics. We present the performance of our algorithm in terms of reduction of the steps between homepage and target. In

\(^2\)http://www.tweepy.org
\(^3\)https://pypi.python.org/pypi/lda
\(^4\)http://www.nltk.org/
\(^5\)https://pypi.python.org/pypi/haralyzer/1.0.7
\(^6\)https://networkx.github.io
\(^7\)http://www.alexa.com/
\(^8\)https://pypi.python.org/pypi/lda
5.3. Experimental Results

Figure 5.3: HTTP archive record (HAR file) extracts the page load, the number of object request and the size in kilobytes of webpage.

addition, we formulate metrics for the complexity of browsing paths.

Distance Improvement Metric

We use the metric to count the reduction of path distance between a random source and a target as studies [4, 84] provides. $AvgD$ is the average length of a collection of minimum length paths as Equation 5.8 presents.

$$AvgD = \frac{\sum_{v \in G-\{root\}} minDist(root, v)}{Total\_Paths} \quad (5.8)$$
5.3. Experimental Results

Path Page Load Metric

We propose a metric to calculate the overall page load of a path as Equation 5.9 presents. We declare that \( \text{PageLoad}_{i,j} \) is the page load in ms of page \( i \) in the path \( j \).

\[
\text{AvgPL} = \frac{\sum_{j \in \text{Path}} \sum_{i \in G} \text{PageLoad}_{i,j}}{\text{Total.Paths}} \quad (5.9)
\]

Path Page Size Metric

The metric of Equation 5.10 reflects the size of browsing memory that a user will need, which has an immediate effect on the browsing experience. We define \( \text{PageSize}_{i,j} \) the page size in kilobytes of page \( i \) in the path \( j \).

\[
\text{AvgPS} = \frac{\sum_{j \in \text{Path}} \sum_{i \in G} \text{PageSize}_{i,j}}{\text{Total.Paths}} \quad (5.10)
\]

Path Object Requests Metric

The path object requests metric of Equation 5.11 calculates the number of objects i.e., images, text, ajax requests, post requests; that is requested from the server by the client. We define \( \text{objects}_{i,j} \) as the number of requested objects of page \( i \) in the path \( j \).

\[
\text{AvgPO} = \frac{\sum_{j \in \text{Path}} \sum_{i \in G} \text{objects}_{i,j}}{\text{Total.Paths}} \quad (5.11)
\]

5.3.3 Results

We discuss our results in the following section. Figure 5.4 presents the average distance improvement calculating paths’ distances from homepage to target nodes between the initial graph \( G \) and our implementation that produces a new graph with personalized hotlinks \( G \cup PH \) using the metric \( \text{AvgD} \) as Equation 5.9 defines. We gradually increase the size of the website’s DAG to analyze the performance. We infer that our implementation improves distance by an average of 11%.

In Figure 5.5, we compare the average path page load in ms, between the initial graph \( G \) and graph derived from our methodology \( G \cup PH \) based on metric \( \text{AvgPL} \). We present the average page load and we gradually increase the size of the website’s DAG. We infer a reduction of time that a user needs for path browsing is about 17.5% as Figure 5.6 provides.
5.3. Experimental Results

Figure 5.4: Improvement % on average path distance between $G$ and $G \cup PH$.

In Figure 5.7, we compare the average number of objects requested on paths in the initial graph and the new graph with personalized hotlinks $G \cup PH$ using the metric $AvgPO$. Also, in Figure 5.8 we present the improvement % on an average number of path requested objects as we gradually increase the size of the website’s DAG. We consider that the new browsing paths request fewer objects. Our implementation reduces the number of object requests by an average of about 14.5%.

In Figure 5.9, we compare the average path page size in Kb, between the initial graph $G$ and the graph produced by our methodology $G \cup PH$. We gradually increase the size of the website’s DAG and we calculate metric $AvgPS$. We consider that our
5.3. Experimental Results

Figure 5.6: Improvement % on average path page load between $G$ and $G \cup PH$.

Implementation reduces the volume of the memory needed by an average of 20% as Figure 5.10 presents.

Also, we analyze the performance in page load, page size and the number of object requests as well as the number of object requests of browsing paths in different webpage categories. We extract the top categories on the website based on the number of their pages/descendants. The topics sports, capital, music, literature, and other constitute the categories. Figures 5.11, 5.12 and 5.13 examine the average page load, page size and number of object requests in the path respectively for the different categories of the website.
5.3. Experimental Results

We infer that the sports category has the most significant improvement on page load and reduces it by about 58%. On page size and on object requests the results are similar; we observe an improvement of about 25% in both dimensions. The capital category is the second in ranking with an average page load improvement of about 31% as well as page size and object requests with an average improvement of about 22%. 

Figure 5.8: Improvement % on average number of path requested objects between $G$ and $G \cup PH$.

Figure 5.9: Average path page size between $G$ and $G \cup PH$. 

We infer that the sports category has the most significant improvement on page load and reduces it by about 58%. On page size and on object requests the results are similar; we observe an improvement of about 25% in both dimensions. The capital category is the second in ranking with an average page load improvement of about 31% as well as page size and object requests with an average improvement of about 22%.
5.4 Concluding Remarks

This chapter deals with personalized advertising inside of a dynamic website. Our innovative methodology enhances the browsing experience of users to reach in fewer steps the necessary information which actually connected with advertising content. We adjust a holistic procedure of website reconstruction in order to place advertising content in an efficient way and improve the performance of the marketing campaign. We use the concept of hotlink assignment to reconstruct the website. A common feature of previous studies is that they deal with the popularity of pages without
5.4. Concluding Remarks

The use of social media contributes to the recalculation of the popularity of each webpage. It is noted that the most of the studies in hotlink in hotlink assignment focus on the number of clicks during browsing from users as a determinant factor to measure the popularity of a webpage. We differentiate at this point using the ranking of trend topics as a factor in the calculation of the popularity.

Many times users feel unsatisfied in the waste of time and thus the reduction taking into consideration information from social media.

Figure 5.12: Improvement % on average path page size in categories between $G$ and $G \cup PH$.

Figure 5.13: Improvement % on average number of path requested objects in categories between $G$ and $G \cup PH$. 
of steps and overall page load constitute significant factors of users satisfaction and improve the consumer’s engagement. Based on our knowledge this study is the first that combines websites’ complexity in terms of page load with hotlink assignment and introduces this criterion in the decision.

Furthermore, we propose a slightly different Pagerank metric i.e. Aux; that takes into account the context similarity and page load of nodes in graph node ranking. The Aux value in the nodes of the graph with hotlinks produces a ranking list of web pages and Ads.

According to our experimental procedure, the proposed methodology enhances the browsing experience in terms of distance and reduces about 11% the steps to reach the webpage target. Also, in our experiments we reduce the time and memory loss on about 17.5% and 20% respectively during the browsing.
5.4. Concluding Remarks
Chapter 6

Pattern Based Diffusion of Marketing Campaign

"First they ignore you, then they laugh at you, then they fight you, then you win."
— Mahatma Gandhi

6.1 Introduction

In a social media marketing campaign, we are interested to identify the initial users that are willing to pass the message to their friends and achieve a maximum diffusion, this is the seeding process. After seeding information within the network, customers’ communication results in information spread. We propose a novel way of identifying the initial first stage actors, using history information of their previous patterns of diffusion. In the current literature, consumer referral behaviour and several antecedents have been proposed for identifying such users. The scarcity of the product, information value, consumer’s need to reciprocate are some examples of antecedents that have been examined as discussed in study [67]. In our approach, we take a different view harnessing available objective data from the social media. We base our seeding policy on users’ behaviour as extracted from a diverse set of recent actual user network participation data.

The information diffusion process that simulates the actual spread of word, has received high attention in the literature mainly using probabilistic models. It has been associated with virus contamination and a plethora of tools and models have been developed trying to explain the spread of information [5, 128]. In addition, there are probabilistic models that have been proposed to capture the maximization of
influence spreading, addressing the problem of identifying the k initial users that will maximize diffusion [62, 63]. Recently the opportunity to harness automatically high volumes of information of different type from social media gives a new dimension to the processing of users’ communication behaviour. In our work, we take a new approach in information diffusion and we propose the effective use of a diverse set of data, i.e., the actual network structure, the behavioural metrics of the user, the text of the actual tweets for a specific topic for developing a realistic policy for a marketing campaign.

We focus on Twitter that has become the most popular microblogging platform, producing every day a vast amount of tweet information. Moreover, we choose to examine information dissemination in Twitter because it enables an effective mechanism of information sharing through the "re-tweet" activity. A retweet is a valuable, interesting and worth sharing information for users of the social media [89] while this type of post depicts the certain adoption of a tweet [90] by the users who made the retweets. Methodologies of prediction and modelling retweet occurrences of a given post have been proposed in order to understand and control information diffusion on Twitter [24]. Thus re-tweets are an explicit way to identify information diffusion. Although information might have been diffused in the network through other activities i.e. relevant tweets, the re-tweet activity is the one that guarantees that there is no error in the sense that the information was diffused as a response to the initially placed message.

Our main contribution is the exploitation of diffusion information patterns, in the selection of nodes that attain the better coverage of the graph. A graph coverage denotes the number of nodes that diffused the information. These nodes are called activated nodes. Our rationale is that based on the availability of a plethora of information that captures a variety of aspects of users’ behaviour, we can develop a methodology that captures the spread of information that can possibly occur in a network in a realistic manner. We propose the Pattern Based Diffusion algorithm attributes information diffusion on users’ history. This allows us to identify the basic diffusion characteristics of the users in a realistic manner. In order to validate our results, we compare our approach with the main existing solutions. Our results show that a marketing campaign that gives incentives to the type of users we propose, can achieve similar results to a probabilistic model of diffusion. Moreover, to further validate our approach we use a methodology to measure the actual network diffusion.

The remainder of the chapter is organized as follows. We provide a detailed description of the methodology in Section 6.2 that we propose which is separated in
sub-modules. We describe in Section 6.2.1 the crawling of the necessary Twitter data. In Section 6.2.3 the predicting model for retweet diffusion is presented. Moreover, in Section 6.2.4 diffusion processes are discussed and our methodology of pattern based diffusion is presented. Our experimental results are presented and discussed in Section 6.4. Finally, in Section 6.5, we present our concluding remarks.

6.2 Methodology

We consider that one of the major tasks in marketing campaigns is the diffusion of specific advertisements or offers to a large number of audience in order to maximize the possibility a part or all of them to be future customers of the product. Due to budget and time limitations, this propagation is not achievable sending a message to all users of the social media but selecting the nodes that can diffuse efficiently the message to their connections. Based on this concept our methodology introduces a pattern based diffusion algorithm which predicts the pattern of diffusion of each user and introduces a propagation procedure similar to message propagation cascades [62, 63].

![System architecture of pattern based diffusion](image_url)

Figure 6.1: System architecture of pattern based diffusion

A summary of the independent tasks that our methodology consists of is given below:

- Crawling Twitter Data. Gathering the data from the online social network which are necessary for the tasks including the tweets, list of followers and
social behavior metrics. Also, creating the graph in a user-oriented Breadth-First search (BFS) approach.

- **User Profiling.** The appropriate information for each user is gathered and a feature vector is created by the sentiment, linguistic, behavior characteristics for each unique user of the dataset.

- **Prediction of Re-tweet Pattern.** A prediction model is trained to categorize the diffusion of a tweet. Each user is categorized to one certain tweet tree diffusion structure.

- **Pattern based diffusion.** Selecting the best nodes for message propagation in terms of time and cover of the social graph.

### 6.2.1 Crawling Twitter Data

The crawling procedure is handled by a dedicated crawler that allows sampling of the Twitter data in a manner that the necessary information for the following task is acquired. The crawling is separated in 2 axes. The first one is the gathering of the scale/range of tweets and the user characteristics that are introduced for the training of the prediction model. The second is the creation of the graph. Instead of using the streaming API, as most existing crawlers do, the proposed crawler performs a Breadth-First search (BFS) starting from one (or multiple nodes) in order to generate the graph. We store the crawled data using a unique identifier for each user account.

### 6.2.2 User Profiling

We consider that the propagation of a tweet is related to the behavior of a user inside of the social media. Also, the type of writing in terms of linguistic format, the content and sentiment of the post constitute significant factors for the reproduction of this post from other users. We extract 2 independent set of features. The first set includes the user communication behavior and the general status in Twitter, i.e., the number of followers, direct tweets, re-tweets, comments and hashtags and the post frequency. The second set is derived from the sentiment and linguistic analysis using the 2015 version of the psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) [117].

The LIWC contains:
6.2. Methodology

- 4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters).

- 22 standard linguistic dimensions (e.g., the percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.).

- 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes).

- 7 personal concern categories (e.g., work, home, leisure activities).

- 3 paralinguistic dimensions (assents, fillers, nonfluencies).

- 12 punctuation categories (periods, commas, etc).

Since LIWC has been proposed, it has been widely used for a number of different tasks, including sentiment analysis [102] and discourse characterization in social media platforms [34].

6.2.3 Prediction of Retweet Pattern

This task focuses on the prediction of user retweet pattern. We treat this concept as a classification problem using machine learning techniques. Our main focus is to train an efficient classifier using the user profile features in order to predict the diffusion pattern of a user. We measure the propagation of a tweet representing it as tree diffusion structure. We define as root node the initial posting of the tweet and each child of the root the retweet from a follower. The tree structure is expanding until the last retweet from a follower be created. Hence each tree structure represents all the followers that retweeted the same message as the root user. According to existing results [76, 73], there are no large cascades of information and the diffusion structure is tend to be in a small scale in terms of depth and width. We introduce 4 patterns based on the experimental results in the crawled data of tweets and their propagation as Figure 6.2 presents. The unique node pattern $L_1$ are users whose tweets are not re-tweeted by followers. The paths of 2 and 3 nodes are derived from users whose tweets are re-tweeted in depth 1 and 2 and labelled $L_1$ and $L_2$ respectively. There is also the case of pattern label $L_4$ that tweets of a user are re-tweeted by 2 different followers creating a split in the diffusion structure. Given the set of most frequent patterns of diffusion for users on Twitter, our next step is to develop a matching process that given a Twitter user it can associate its pattern of diffusion. We annotate the diffusion
6.2. Methodology

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.88</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.71</td>
</tr>
<tr>
<td>SVM</td>
<td>0.77</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.61</td>
</tr>
</tbody>
</table>

structure of each user based on the similarity with the 4 patterns. We represent each user as a vector of features and a label of tree diffusion structure.

The feature vector is introduced in the learning process of classifiers. The most efficient classifier in terms of performance is used. We separate the dataset to training and test set, using the K-Fold Cross-Validation with \( K = 10 \). In Table 6.1, the comparative study is presented. We consider that the examined classifiers achieve an F-Measure above 60% and the best one is Random Forest with \( F1 = 0.88 \). The accuracy of our prediction is high which validates that based on the linguistic aspects of the tweet and the behavior status, we can predict the tree diffusion structure.

Our approach considers the fact that in due time the user might change behaviour hence the classifier can be trained with a new dataset that represents the users’ updated feature vector. Our current work utilizes this prediction of tweet patterns which is based on user behaviour, for assigning to each user in the Twitter graph his/her predicted tweet pattern behaviour.

![Figure 6.2: Tweet Tree Diffusion Structure Patterns](image)

6.2.4 Pattern Based Diffusion

The influence maximization problem is a problem that has been discussed extensively in the literature and is defined as finding \( k \) individuals that can result in the largest
diffusion of information within the network [62, 63]. In the literature, there are two basic models: the Independent Cascade model (IC) and the Linear Threshold model (LT). In both models, at any time step, a user is either active (an adopter of the product) or inactive. In the IC model, each edge has an activation probability and influence is propagated by activated nodes independently activating their inactive neighbours based on the edge activation probabilities. When an inactive user becomes active at a time step $t$, it gets exactly one chance to independently activate its currently inactive neighbours at time $t + 1$. In the LT model, each edge has a weight, each vertex has a threshold chosen uniformly at random. The sum of incoming edge weights on any node is assumed to be at most 1 and every user chooses an activation threshold uniformly at random from $[0, 1]$. At any time step, if the sum of incoming influence (edge weights) from the active neighbours of an inactive user exceeds its threshold, it becomes active. In both models, influence propagates until no more users can become active.

Our approach to network information diffusion is based on an idea that considers users’ propagation patterns. We claim that when taking into account each user’s specific tendency in information diffusion, a more realistic information diffusion process can be constructed as compared to IC and LT where no previous knowledge is considered. Our objective is to come up with a marketing campaign policy that can achieve information spread similar to the best of IC and LT in a realistic manner, activating the right nodes. Hence we are not presenting yet another algorithm for information diffusion, but a process that can actually achieve that diffusion. Our proposed Pattern Based Diffusion algorithm (PBD) is an approach that examines how information is diffused within the network when a different type of users are triggered. The input of the PBD algorithm is the social network where each node is enriched with its pattern of diffusion. Similarly to other diffusion processes, we take a greedy approach selecting at a given point in time to activate those users that adhere to a specific pattern or patterns.

Given the social graph and the preferable pattern, i.e. $L_2$, the PBD is examining whether there are any neighbors of the given pattern and if yes one $L_2$ node is activated. If there is no neighbor node of $L_2$ then based on a priority the other patterns are examined. In case there are more than one neighbor nodes for $L_2$ then the procedure Resolveties is invoked which selects the neighbor node whose children have the greatest degree. If more than one exists then a random selection is made. In case there is no $L_2$ neighbor node, then the other patterns are examined next($L_i$). The examination order is $L_2$-$L_3$-$L_4$. The rational of the greedy algorithm is given a
preferable pattern, to continue the diffusion based on this pattern and if this is not possible using the pattern that spreads the information more. In the case of $L3$, not only the neighbor but the child of the neighbor is activated as well. In the case of the $L4$ two neighbor nodes are selected instead of one.

There different variations of the PBD algorithm, each one uses a different pattern or pattern combination as the preferable pattern for diffusion. Our objective is to find out which is the diffusion pattern that can achieve a diffusion that is similar to the one simulated by IC or LT. The use of the PBD algorithms provides a more realistic information diffusion process if a certain pattern of diffusion is adopted. We use the PBD to identify the pattern of the k initial nodes that can lead to an information diffusion that can reach the outcome of IC.

6.3 Estimating the Actual Propagation

As already mentioned one of the major issues in the discussion of information diffusion models is how to verify the validity of the model. There are several restrictions that make the problem of actually asserting the extent of information diffusion a difficult one to define. Users might not discuss on the same topic, but they might discuss on related issues. A user might not re-tweet a specific message, but he/she can start discussing it after reading a friend’s tweet. In our study, we propose a new alternative in identifying the actual diffusion where the propagation of the message happened when a certain message-tweet belongs to a high-quality dense cluster of messages-tweets in a specific topic. We propose a methodology for evaluating the actual diffusion within a Twitter network, based on similarity of users’ tweets. Within the network, each user has an associated set of tweets. Information diffusion takes place through these tweets. We aggregate tweets by users and in this way we formulate longer text bursts. We measure network diffusion based on a users’ tweets similarity measure. Our rationale is that on a topic based sampling, users’ tweets that show great similarity are referred to the same or similar topic. For every user we extract his information vector, using linguistic analysis. The most important aspects of user’s tweets are represented as features within a vector. We view these tweets as points in the space and we identify a tweet cluster of high quality. Such a dense cluster of users whose tweets are similar imply that users are discussing similar topics. Our argument is based on the fact that tweets with the same words in their feature vector are close. We use a pre-filtering clustering step with k-means and locate these high-quality clusters. The objective of the pre-filtering step is to identify the number of k
Algorithm 6.1 Patterns Based Diffusion

1: **input** social graph, \( l_i \)
2: **output** List of marked nodes
3: \{\( l_i \) is the selected priority pattern\}
4: \( N = [n_1, ... n_k] \)
5: \( M = [n_1, ..., n_k] \) \{M is the output, the list of the marked nodes\}
6: \( Tested = \emptyset \) \{Tested is the set that holds all the neighbor nodes that are candidates for diffusion\}
7: \( New = \emptyset \) \{New is the array that holds the new marked nodes to be examined for diffusion.\}
8: Select \( k \) initial nodes based on \( l_i \)
9: **while** \( N \) is not empty **do**
10:  **for each** \( n_i \in N \) **do**
11:      \textit{inserted} == false;
12:      **while** \( \text{inserted} == false \) \&\& \( \text{count} < 2 \) **do**
13:         **for each** \( n_{ij} \in \text{Neighbors}(n_i) \) **do**
14:            if \( n_{ij} == \text{unmarked} \) \&\& \( \text{pattern}(n_{ij}) == l_i \) then
15:               \( Tested = Tested \cup n_{ij}; \text{count} += 1 \)\end{algorithm}
16:          end if
17:          if \( Li == (L2 \| L3) \) then
18:             \textit{inserted} == true
19:          else if \( Li == L4 \) \&\& \( \text{count} == 2 \) then
20:             \textit{inserted} == true;
21:          end if
22:      end for
23:      \( \text{pattern} = \text{next}(L_i) \); \{since no priority pattern was found, other patterns are searched\}
24:  **end while\}
25: **if** \( \text{size}(Tested) > 1 \) **then**
26:  \( n_{ijk} == \text{Resolveties}(Tested) \)
27: **end if**
28: **if** \( Li == L3 \) **then**
29:  \( M = M \cup \text{father}(n_{ijk}); New = New \cup \text{father}(n_{ijk}); \)
30: **end if**
31: **end for**
32: \( N = New; New == \emptyset; \text{marked} == \emptyset \)
33: **end while**
6.4. Experimental Results

6.4.1 Implementation

We based our experiments on Twitter and used Twitter API to collect tweets. We contracted our experiments using data that we collected from the public Twitter API using the tweepy\(^1\) tool. To implement the various network related feature extraction methods we used networkx\(^2\). We used the ML algorithms provided by scikit-learn\(^3\) to train all classifiers The IC, LT as well as PBD are implemented in Python 2.7. The crawling process is topic-centric. We used a dataset of tweets that were published for a time interval of 21 days using the keyword #MH370 concerning Malaysia Airlines Flight 370 disappearance. We used the resulted dataset to extract users that have discussed the incidents, their related tweets and the tweets of their followers in-depth two. For the user pattern creation, we extracted the last 1000 tweets of each user and

\(^1\)http://www.tweepy.org  
\(^2\)https://networkx.github.io  
\(^3\)http://scikit-learn.org
examined the retweets. We crawled the Twitter using the above keyword a dataset was created containing 13000 Tweets that had been done by 11130 users. We categorized each user in a certain pattern converting each structure to a string and comparing by edit distance the real diffusion structure with the fixed categories of diffusion as Figure 6.2 presents. The tweets for each user was analysed with LIWC software and thus linguistic and emotional characteristics were extracted. In addition, features derived from the social behavior of each user were added in the vector. The training of a Random Forest classifier was implemented in order to predict the category of diffusion structure. Our results in Table 6.1 show that the Random Forest classifier achieves the highest F-Measure value (0.88) for this type of data.

### 6.4.2 Results

In our experimental tries, we implemented PBD and executed experiments for different labels. We compared our results with the outcome of IC and LT. In Figure 6.3 we observe that the IC algorithm is producing the maximum cover 18% of the graph within less than 10 steps. After these 10 steps, no more nodes are activated. The LT achieves less cover reaching the 12% in less steps than IC. We executed as shown in Figure 6.3 the variations of PBD as well. PBD diffuses information giving priority of activation to a specific node pattern while in the random version a random pattern is selected.

All heuristics have been simulated for 10000 rounds, in the same manner as in [62]. Figure 6.3 shows the average values for each algorithm that was executed 10000 times. Our results show that if a PBD with priority on the L3 pattern is selected then a cover similar to the one produced by IC can be achieved. Also, we observe that our approach achieves results similar to IC, but using more steps. This is an expected result since in a realistic environment information propagation occurs gradually and our PBD algorithm tries not only to find the set of active nodes but to simulate the process in a realistic step by step manner.

Figures 6.4(a)-6.4(f) show for each algorithm the cover for the 10000 rounds where in each round a separate initial set of k nodes was selected based on the criteria set by the given algorithm. It is interesting to notice the different behavior of PBD variations and the rest approaches. We observe that within the next almost 30 steps L3 and L2 achieve similar cover. Hence for the initial steps, all executions behave similarity because while nodes are available the choices are given. When more of the 12% of the graph is covered then variations occur based on the availability of routes. Still, in L3 for example, the variations result in executions that achieve in the end
6.4. Experimental Results

Figure 6.3: The average rate of coverage. The x-axis presents the number of steps and y-axis the average rate of coverage based on 10000 rounds of simulation.

16 to 20% cover. On the other hand, the random version shows a higher range of graph cover achieved in each step as well as in the final case. The same holds for IC. These results show that in our approach where a specific type of nodes is triggered for information propagation it can achieve in a more consistent manner the calculated cover.

As a next step for the verification of our methodology, we calculated the actual diffusion. Our objective is to identify a quality cluster of similar tweets. This cluster determines the information diffusion. As shown in Figure 6.5(a), we experimented from 2 until 12 different clusters and we computed the average silhouette value. From our results, it is evident that the best quality clustering is achieved for four clusters. Figure 6.5(b) shows the computed silhouette value for k equals to 4. It is evident that there is one dense and high-quality graph, produced by cluster 2 where most of the silhouette values are near 0.8. This cluster achieves a 19% cover of the graph i.e 19% if the nodes are activated (Table 6.2). The results on clustering in real data show the high-quality dense cluster reaches the desirable diffusion of IC and PBD. We extracted all nodes that belong to graph one and we calculated that they represent the 18% of the graph. Hence based on the actual verification of the information spread in the graph the 18% was activated. These results show that the IC for \( p = 0.1 \) and our PBD/L3 can achieve the actual diffusion.
6.4. Experimental Results

Figure 6.4: Results of each heuristic simulation. The x-axis presents the number of steps and y-axis the average rate of coverage based on 10000 rounds of simulation.

6.4.3 Discussion

Our methodology can help marketers to implement an effective marketing campaign. Given the current diffusion theories, using probabilistic models the algorithms can
6.4. Experimental Results

(a) The average silhouette values increasing the number of clusters

(b) Silhouette values in $k=4$

Figure 6.5: Quality of clustering

Table 6.2: Results of clustering $k = 4$

<table>
<thead>
<tr>
<th>Clustering k-means</th>
<th>% of graph cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0.004</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.19</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.69</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.10</td>
</tr>
</tbody>
</table>

estimate the diffusion of the information. But when it comes to influencing this diffusion not much information is available. Our methodology proposes a way to achieve a diffusion rate as the one predicted by the IC.

We discover the $k$ nodes of the graph that can maximize the diffusion given a certain type of message. The real data was used in order to identify the diffusion pattern for each node using a classifier. We compare well-known probabilistic models IC and LT with our algorithm which is more realistic. Also, the IC is based on a probabilistic factor in each edge and thus we examine in which case this factor reaches the same coverage of nodes as our algorithm.

We introduce a general framework for identifying the diffusion patterns. The prediction of diffusion patterns is based on the linguistics and social behavior metrics in a time window that they will be gathered and thus the current changes in the diffusion pattern maybe are not visible. However, the feature vector can be feed with current tweets which it will lead to the current diffusion pattern. The feature vector has been created based on the linguistic and emotional characteristics derived from LIWC [95] software so as tweets with similar semantic and sentiment will be close to a multidimensional space. The prediction of diffusion pattern for each user is
6.5 Concluding Remarks

We looked into the problem of finding the pattern behaviour of the users that can achieve an information diffusion result similar to the expected actual diffusion achieved by word of mouth. Our methodology is based on the usage of diffusion patterns that have already been registered in a specific social graph. We have verified our methodology performing a set of experiments that show that our proposed PBD algorithm can activate a social network in a similar way to the IC algorithm. Moreover, we validated the spread of information, proposing a methodology to calculate the actual diffusion based on tweets’ clustering.

As a future work, we are considering to explore the combination of PBD with probabilistic models and experimenting in creating different heuristics based on the
6.5. Concluding Remarks

Figure 6.6: Independent cascade model

LT or IC. Furthermore, our technique could be expanded using a community detection algorithm so as to choose the best k-nodes from the representatives of each community. We are interested to investigate a parallel implementation of our methodology applied simultaneously in different communities of the social graph and merging the results to the whole graph in order to achieve better performance.
Chapter 7

Concluding Discussion and Future Work

"The present is theirs; the future, for which I really worked, is mine."
— Nikola Tesla

This dissertation focuses on the development of innovative data mining techniques for the purpose of effective marketing campaigns. Our work contributes to the exploration of new methodologies in marketing campaign targeting, conversational and personalized advertising as well as the information propagation in Online Social Networks.

Regarding marketing campaign targeting, we present an innovative methodology to improve the efficiency of marketing campaigns through the identification of nodes with high bridge participation in the evolving social network. We provide the detailed description of the methodology for predicting the future linkage of users in the social network, which constitutes the basic component of our model. We recommend an alternative metric for analyzing user influence, the bridge participation in the evolving graph. We infer that the combination of topological, behavioral and linguistic features of multilayer Online Social Networks reach the highest prediction performance and the XGBoost classifier perform more efficiently the link prediction based on our empirical experimental study. Also, our procedure proposes the automated recognition and removal of the removal of users that constitute content polluters or bots from the set of candidate targets. As future work, we are interested in developing our methodology in heterogeneous social networks such as Facebook and LinkedIn and recognize the factors that influence the outcomes of our algorithms in a finer granularity level. In addition, we consider that location oriented features could improve the performance of the link predictor. Furthermore, our current interests are directed on utilizing cloud
computing technologies in our method and the study of scalability issues dealing with link prediction and efficient marketing campaigns.

We present the experimental results of social bots infiltration and how these accounts can be involved in the process of automated promotion and conversational advertising. We introduce dedicated social bots with AI chat functionality that can interact with other users on Twitter. We infer that the ability to chat does not affect negatively the performance of a social bot contrary to existing behavior-based detection strategies which suggest the limitation or social bots interactions. We focus on the analysis of influence metrics such as Klout, message interactions and number of followers. We result that communication skills create higher influence and infiltration in Online Social Networks than previous studies with social bots having basic characteristics. As future work, we are interested in examining the infiltration strategies of social bots in different social networks such as Facebook and Instagram.

We provide an efficient methodology for placing of advertising content inside a dynamic website. We combine the concept of hotlink assignment with personalization through social media and we adjust an automated reconstruction of the website that places advertisements in the browsing paths of users. We use a generic methodology for personalization which captures users’ posts. The additional links, i.e., hotlinks enhance the browsing experience of users by reducing the steps that users do to reach a popular web page of their interests. The proposed algorithm of personalized hotlink assignment takes into consideration popularity of the web page, the context similarity, the trend topics of social media and the complexity of a web page in terms of page load. Furthermore, we propose a slightly modified PageRank metric, i.e., Aux; that calculates the ranking using the context similarity and page complexity. The Aux value in the nodes of the personalized graph with hotlinks produces a ranking list of web pages and advertisements. In future work, we are interested in the adaptation of the proposed methodology in different types of marketing campaigns, e.g., elections, brand promotions, job offers. Also, we would like to investigate how other social networks influence the results of our algorithm.

Related to information diffusion in social media, we looked into the problem of finding the pattern behaviour of the users that can achieve an information diffusion result similar to the expected actual diffusion achieved by word of mouth. We introduce a novel methodology to achieve information diffusion within a social graph that activates a realistic number of users. Our approach combines the predicted patterns of diffusion for each node with propagation heuristics in order to achieve an effective cover of the graph. The novelty of our methodology is based on the use of history
information to predict users’ diffusion patterns and on our proposed heuristics for achieving a realistic information spread. Our methodology is based on the usage of diffusion patterns that have already been registered in a specific social graph. We verify our methodology performing a set of experiments that show that our proposed Pattern Based Diffusion algorithm can activate a social network in a similar way to probabilistic models in terms of cover rate. Moreover, we validate the spread of information, proposing a methodology to calculate the actual diffusion based on tweets’ clustering. As a future work, we are considering to explore the combination of PBD with probabilistic models and experimenting in creating different heuristics based on the LT or IC. Furthermore, our technique could be expanded using a community detection algorithm so as to choose the best k-nodes from the representatives of each community. We are interested to investigate a parallel implementation of our methodology applied simultaneously in different communities of the social graph and merging the results to the whole graph in order to achieve better performance.
Appendix A

Summary in Greek

Α’.1 Εισαγωγή

Η δημοτικότητα των Κοινωνικών Δικτύων αυξάνεται καθημερινά, όπως δείχνουν τα τρέχοντα στατιστικά στοιχεία. Συγκεκριμένα, το τρίτο τρίμηνο του 2017, το Facebook και το Twitter έχουν 2.07 δισεκατομμύρια 1 και 330 εκατομμύρια 2 μηνιαίους ενεργούς χρήστες αντίστοιχα. Οι άνθρωποι χρησιμοποιούν συνεχώς τις πλατφόρμες κοινωνικής δικτύωσης για να μοιράζονται περιεχόμενο σχετικά με την καθημερινότητά τους και τα πράγματα που συμβαίνουν γύρω τους. Αυτά τα συστήματα έχουν φέρει επανάσταση στον τρόπο με τον οποίο επικοινωνούμε οργανώνοντας τις κοινωνικές σχέσεις μας σε ψηφιακή μορφή. Ταυτόχρονα, τα μέσα κοινωνικής δικτύωσης εξυπηρετούν τις ανάγκες των ατόμων που ασχολούνται με εκστρατείες προώθησης προϊόντων και διαφημιστικών μηνυμάτων εξαιτίας της τεράστιας συμμετοχής σε αυτές τις πλατφόρμες και των μεγάλων δυνατοτήτων βελτίωσης των στρατηγικών τους. Η επιλογή των στόχων, η διάδοση των μηνυμάτων προώθησης εμπορικών σημάτων και η θέση του διαφημιστικού περιεχομένου σε δημοφιλείς ιστοσελίδες αποτελούν μερικούς από τους στόχους εκστρατειών, προκειμένου να αποκτήσουν το ενδιαφέρον των δυνητικών πελατών. Αυτή η τεράστια δημοτικότητα των κοινωνικών μέσων όχι μόνο παρέχει την ευκαιρία της ανίχνευσης χρήστες και διασυνδέσεις τους, αλλά και της εφαρμογής αλγορίθμων και τεχνικών εξόρυξης δεδομένων, προκειμένου να δημιουργηθούν χρήσιμα εργαλεία και αποτελεσματικές στρατηγικές σε διαφημιστικές εκστρατείες. Οι τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης στον τομέα του μάρκετινγκ έχουν κερδίσει το ενδιαφέρον των ερευνητών για την ανάπτυξη εργαλείων

α’.2. Κίνητρο

υπολογισμού της επιρροής των χρηστών, την αναγνώριση των αυτοματοποιημένων μη-
χανών, την διάχυση πληροφοριών, την εξαγωγή του ενδιαφέροντος των χρηστών κτλ.
Αυτή η διατριβή επικεντρώνεται σε αυτό το πλαίσιο. Στόχος μας είναι να σχολιάσουμε με
συγκεκριμένες εφαρμογές εκστρατειών μάρκετινγκ χρησιμοποιώντας τεχνικές εξόρυξης
dεδομένων. Μας ενδιαφέρει η διερεύνηση νέων μεθοδολογιών στην στόχευση συγκε-
κριμένων χρηστών, η εξατομικευμένη διαφήμιση καθώς και η διάδοση πληροφοριών στα
κοινωνικά δίκτυα. Ελπίζουμε οι μεθοδολογίες μας να είναι χρήσιμες στα συστήματα συ-
στάσεων καθώς και στις επιχειρήσεις που ενδιαφέρονται να χρησιμοποιήσουν ανεξάρτητα
tις προτεινόμενες τεχνικές μας ή ως ενιαίο πλαίσιο για την εκτέλεση αποτελεσματικών
eκστρατειών μάρκετινγκ

Α’.2 Κίνητρο

Σήμερα, τα διαθέσιμα δεδομένα από τις πλατφόρμες κοινωνικής δικτύωσης δίνουν μια μο-
ναδική ευκαιρία να εξάγονται μοτίβα που σχετίζονται με τα ενδιαφέροντα των χρηστών,
tις διασυνδέσεις τους και το επίπεδο επιρροής τους σε άλλους χρήστες. Θεωρούμε ότι
λόγω των περιορισμών του προϋπολογισμού και της χρονικής διάρκειας, οι εταιρείες που
eνδιαφέρονται για την προώθηση επικεντρώνουν μόνο σε ένα συγκεκριμένο σύνολο
χρηστών (target group) επιδιώκοντας τη βελτιστοποίηση της εκστρατείας τους. Παρά
τη πληθώρα των εργαλείων συλλογής και ανάλυσης δεδομένων, η διαδικασία στόχευσης
περιορίζεται στην επιλογή ενός συνόλου χρηστών, που ονομάζονται χρήστες επιρροής
(influential) με βάση των αριθμών των συνδέσεων ή την αναπαραγωγή των αναρτήσε-
ων τους. Ωστόσο το κοινωνικό δίκτυο αποτελεί ένα εξελιγμένο φαινόμενο και έτσι
η συμπεριφορά και ο αντίκτυπος των χρηστών αλλάζουν κατά τη διάρκεια του χρόνου.
Θεωρούμε ότι αυτές οι πληροφορίες διαφοροποιούν τη λήψη αποφάσεων κατά τη διάρκεια
μιας εκστρατείας μάρκετινγκ. Οι τρέχουσες προσπάθειες δεν έχουν αυτή τη διάσταση
που μας ωθεί να ερευνήσουμε νέες μετρικές. Ένας τρόπος για να μοντελοποιήσουμε ένα
κοινωνικό δίκτυο είναι το γράφημα στο οποίο οι κόμβοι και οι ακμές αντιστοιχιζόμενοι
cοινωνικών δικτύων που επεκτείνουν τα τοπολογικά, συμπεριφορικά και γλωσσικά
χαρακτηριστικά θα μας δώσει τη δυνατότητα να αναπτύξουμε αποτελεσματικά εργα-
λεία τεχνητής νοημοσύνης για τη συμβουλευτική υποστήριξη κατά τη διάρκεια μιας
α.3 Στόχοι

Επίσης, τα διαθέσιμα εργαλεία και μεθοδολογίες στερούνται αυτοματοποιημένης προώθησής και οι περισσότερες από τις προσπάθειες χρησιμοποιήσης των κοινωνικών μέσων στο μάρκετινγκ είναι χειροκίνητες. Παρατηρείται ότι η χρήση αυτοματοποιημένων μηχανών ΑΙ είναι κατάλληλη για τη βελτίωση του μαζικού χρησιμού της προώθησης του περιεχομένου και για την άμεση διαδικασία προώθησης των διαφημίσεων εξιδικευμένα στις ανάγκες των χρηστών.

Επιπλέον, οι εταιρείες τείνουν να τοποθετούν το περιεχόμενο των διαφημίσεων σε δημοφιλείς ιστοσελίδες, ωστόσο, οι χρήστες ενοχλούνται από την παρουσία διαφημιστικού περιεχομένου κατά την περιήγησής τους με βάση το γεγονός ότι οι περισσότερες από αυτές δεν έχουν σχέση με τις προτιμήσεις τους. Οι διαθέσιμες πληροφορίες από τα κοινωνικά μέσα παρέχουν τη δυνατότητα εξαγωγής των προτιμήσεων των χρηστών με άμεσο ή έμμεσο τρόπο. Θεωρούμε ότι η βελτίωση της βελτίωσης των χρηστών κατά την περιήγηση συσχετίζεται με την πρόθεση των να λαμβάνουν διαφημιστικό περιεχόμενο ή να αγοράζουν συγκεκριμένα προϊόντα. Η έλλειψη μιας αλητικής προσέγγισης ανακατασκευής ιστότοπου που λαμβάνει υπόψη την βελτίωση της περιήγησης (π.χ. μείωση των βημάτων περιήγησης), των ενδιαφερόντων του χρήστη και την τοποθέτηση θέσης των εξατομικευμένων διαφημίσεων μας παρακινεί σε μια μεθοδολογία για το σκοπό αυτό. Μια άλλη πτυχή είναι η διαδικασία της εξάπλωσης μηχανών, δηλ. του διαφημιστικού περιεχομένου στο κοινωνικό δίκτυο. Η διαδικασία διάδοσης πληροφοριών που προσομοιώνει την πραγματική εξάπλωση, έχει κερδίσει το ερευνητικό ενδιαφέρον χρησιμοποιώντας κυρίως πιθανοτικά μοντέλα. Μια πληθώρα εργαλείων και μοντέλων έχουν αιτιολογηθεί η προσπαθώντας να διερευνήσουν την εξάπλωση πληροφοριών, π.χ. μάλλον από ιστούς. Επιπλέον, υπάρχουν πιθανοτικά μοντέλα που έχουν εισαχθεί για να συλλέξουν τη μεγιστοποίηση της εξάπλωσης επιρροής, αντιμετωπίζοντας το πρόβλημα της αναγνώρισης των αρχικών χρηστών που θα μεγιστοποιήσουν τη διάδοση. Ωστόσο, τα εισαγόμενα μοντέλα δεν έχουν το χρεισμό αυτής της διαδικασίας βασισμένα στην πραγματική διάχυση. Εισάγουμε μια νέα προσέγγιση στην διάχυση πληροφοριών με την αποτελεσματική χρήση ενός ποικίλου συνόλου δεδομένων, δηλ. της τοπολογίας του δικτύου, των μετρικών συμπεριφοράς του χρήστη, των γλωσσολογικών χαρακτηριστικών των αναφερόμενων κειμένων με σκοπό την ανάπτυξη μιας πιο ρεαλιστικής διάδοση των συνέχειαν έρευνα στο κοινωνικό δίκτυο.

Α'.3 Στόχοι

Ο κύριος στόχος αυτής της διατριβής είναι η ανάπτυξη τεχνικών εξόρυξης δεδομένων και η δημιουργία εργαλείων ΑΙ για τη βελτίωση των πολιτικών λήψης αποφάσεων σε
α΄.4. Συνεισφορά

εκστρατείες μάρκετινγκ. Περιγράφουμε τους στόχους μας σε 4 άξονες. Ο πρώτος αποτελεί την πρόταση μιας εναλλακτικής μετρικής επιρροής που λαμβάνει υπόψη την εξέλιξη του κοινωνικού δικτύου και την αλλαγή της κατάστασης των χρηστών. Ο δεύτερος είναι η ανάλυση της διείσδυσης των αυτοματοποιημένων μηχανών AI (social bots) και η χρήση τους στην βελτίωση διαφημιστικών εκστρατειών. Ο τρίτος είναι η πρόταση μιας μεθοδολογίας για την αυτοματοποιημένη τοποθέτηση διαφημιστικού περιεχομένου σε έναν δυναμικό ιστότοπο χρησιμοποιώντας πληροφορίες από κοινωνικά μέσα και παρέχοντας βελτίωση στην περιήγηση. Ο τέταρτος άξονας περιλαμβάνει τον ορισμό των αλγορίθμων που εισάγουμε στην προσομοίωση της διάχυσης των πληροφοριών στο κοινωνικό δίκτυο με πιο φιλικό και εύκολα προσδιορισμένο τρόπο χρησιμοποιώντας τα ιστορικά πρότυπα διάδοσης των αναρτήσεων των χρηστών.

Α´.4 Συνεισφορά

Τα βασικά σημεία της συνεισφοράς μας μπορούν να συνοψιστούν στις ακόλουθες προτάσεις:

- Προτείνουμε μια εναλλακτική προσέγγιση για τον υπολογισμό της επιρροής του χρήστη, τη συμμετοχή των χρηστών σε γέφυρες μεταξύ κοινοτήτων του εξελισσόμενου γραφήματος. Λαμβάνοντας υπόψη τη μετρική επιρροής καθώς και την αναγνώριση των αυτοματοποιημένων μηχανών μέσα στο δίκτυο και εισάγουμε μια σταθμισμένη στρατηγική κατανομής κόστους για την εκτέλεση αποτελεσματικών εκστρατειών μάρκετινγκ.

- Αναλύουμε τη βελτίωση της διείσδυσης των social bots χρησιμοποιώντας αυτοματοποιημένες επικοινωνιακές δεξιότητες και εισάγουμε τα Conversational Social Bots ως προωθητές περιεχομένου διαφημίσεων.

- Παρουσιάζουμε μια αποτελεσματική διαδικασία για την προσομοίωση της διάχυσης σε έναν ιστότοπο. Επίσης, αναπτύσσουμε μια ολιστική διαδικασία ανακατασκευής ιστότοπου χρησιμοποιώντας ένα γενικό σχήμα για εξατομίκευση μέσω των κοινωνικών μέσων και έναν προσαρμοσμένο αλγόριθμο αντιστοίχισης επιπρόσθετων συνδέσμων.

- Προτείνουμε μια νέα μεθοδολογία για τη διάδοση των πληροφοριών που μπορεί να συμβεί σε ένα δίκτυο με πιο ακριβή τρόπο. Εισάγουμε τον αλγόριθμο διάχυσης βασισμένο σε μοτίβα, ο οποίος διερεύνα τη διάδοση της πληροφορίας με βάση τα ιστορικά των χρηστών.
Α’.5 Στοχοποίηση χρηστών

Στην προσέγγισή μας, συνδυάζουμε τοπολογικά, γλωσσικά, συμπεριφορικά και πολυεπίπεδα χαρακτηριστικά για να επιτύχουμε καλύτερες επιδόσεις στην πρόβλεψη συνδέσμων. Στόχος μας είναι να προσδιορίσουμε τους χρήστες που αποτελούν τους σπόρους (χρήστες επιρροής) για καμπάνιες μάρκετινγκ εξετάζοντας τα ποσοστά συμμετοχής σε γέφυρες (υπάρχουσες + προβλεπόμενες) μεταξύ των κοινοτήτων του κοινωνικού δικτύου. Η μεθοδολογία μας όχι μόνο εντοπίζει δυνητικούς σημαντικούς χρήστες, γεγονός που είναι επωφελόμενη για την πρόωθηση συνδέσμων, αλλά επίσης αξιολογεί τον αντίκτυπο των διαφορετικών ποσοστών συμμετοχής για την πρόβλεψη συνδέσμων. Η μεθοδολογία μας όχι μόνο εντοπίζει δυνητικούς σημαντικούς χρήστες, γεγο

Χρησιμοποιούμε παρόμοια μεθοδολογία όπως παρουσιάζεται στην μελέτη [55], η οποία προτείνει τη συμμετοχή των κόμβων σε γέφυρες χοιριντών ως μέτρο επιρροής. Το μοντέλο παρουσιάζεται στο Σχήμα A’.1. Όσον αφορά τη διάδοση των μηχανών, απαιτείται την επιμόνωση των προβλεπόμενων κοινωνικών δικτύων κατανόηση να διαχειριστούν αυτοματοποιημένα μηχανές ΑΙ, αφαιρώντας αυτούς τους λογαριασμούς από τους υποψήφιους στόχους. Η μεθοδολογία μας προτείνει μια σχεδόν μηχανική διακόπτης, με την ιδέα να αξιολογηθεί η ακρίβεια του οποίου και την αξιολόγηση της XGBoost με αποτέλεσμα να καταλήξει σε γέφυρα χοιριντών. Το μοντέλο παρουσιάζεται στο Σχήμα A’.1.
Σχήμα Α’.1: Αρχιτεκτονική συστήματος: Οι ανεξάρτητες λειτουργίες που περιλαμβάνει η μεθοδολογία μας ξεκινώντας από τη διαδικασία δημιουργίας του γραφήματος και καταλήγοντας στη στόχευση της καμπάνιας.

είναι η εκπαίδευση και ο έλεγχος της απόδοσης ενός ταξινομητή για την πρόβλεψη της σύνδεσης κόμβων σε ένα γράφημα που αλλάζει στο χρόνο.

Μια σύνοψη των ανεξάρτητων εργασιών που περιλαμβάνει η μεθοδολογία μας δίνεται παρακάτω:

- **Δημιουργία γραφήματος.** Ξεκινώντας από μια λίστα με κόμβους (σπόρους) από διαδικτυακά κοινωνικά δίκτυα, συγκεντρώνουμε δεδομένα σχετικά με το δίκτυο που εστιάζει στον εαυτό μας για κάθε χρήστη και ανακαλύπτουμε νέους κόμβους σε μια προσέγγιση Breadth-First search (BFS) γράφημα πολλαπλών στρώσεων. Επίσης, ανιχνεύουμε τις πρόσφατες δοκιμές / αναρτήσεις κάθε χρήστη.

- **Εξαγωγή χαρακτηριστικών.** Χειρίζοντας τα κείμενα του κάθε χρήστη, εξάγουμε τα συνασπίζοντα, τα γλωσσικά, τα δημοφιλή και τα διαστρωματικά χαρακτηριστικά για κάθε χρήστη του υπάρχοντος γραφήματος. Επίσης, εμείς οι δομικές

- **Εξαγωγή γεφυρών.** Χρησιμοποιώντας ένα μοντέλο ταξινόμησης για την πρόβλεψη συζύγων μεταξύ των ζευγών κόμβων, τα οποία εκπαιδεύουμε χρησιμοποιώντας τα δεδομένα ανίχνευσης, ανακαλύπτουμε την εξέλιξη του γραφήματος από την αποφυγή των νέων συνδέσμων. Η επέκταση του γραφήματος με τους νέους προβλεπόμενους συνδέσμους δημιουργεί νέες κοινότητες και εξάγονται γέφυρες.
α.5. Στοχοποίηση χρηστών

- Ανίχνευση αυτοματοποιημένων μηχανών. Οι λογαριασμοί που αποτελούν αυτοματοποιημένες μηχανές AI ανιχνεύονται και αφαιρούνται από τη διαδικασία επιλογής των υποψηφίων στόχων.

- Επιλογή στόχων. Οι κόμβοι κατατάσσονται με τη συμμετοχή της γέφυρας. Οι κόμβοι με υψηλή συμμετοχή χαρακτηρίζονται ως σημαντικοί και ενδιάμεσοι μεταξύ των κοινοτήτων, εξάπλωσης μηνυμάτων σε χωριστά τμήματα του γραφήματος.

Σύμφωνα με την πειραματική μας διαδικασία, ο ταξινομητής XGBoost έχει την καλύτερη απόδοση για την πρόβλεψη νέων συνδέσεων στα γραφήματα. Στην περίπτωσή μας, το χρονικό παράθυρο αποτελεί την περίοδο ενός μηνός. Χρησιμοποιήσαμε το εκπαιδευμένο μοντέλο του ταξινομητή XGBoost για να εξετάζουμε τις προβλέψεις συνδέσεων μεταξύ των ζευγών κόμβων σε 2 κοινότητες με τον μεγαλύτερο αριθμό κόμβων.

Στην περίπτωσή μας, το γράφημα mutual-follow για να εκμεταλλευτούμε τους ισχυρότερους δεσμούς μεταξύ των κόμβων.

Η επιλογή μας υποστηρίζεται επίσης από τα πειραματικά αποτελέσματά μας που υποδηλώνουν ότι μπορεί να επιτευχθεί υψηλότερη απόδοση πρόβλεψης όταν αξιοποιείται η αμοιβαία ακολουθούμενη τοπολογία. Ως τελευταίο βήμα, τρέχουμε τον αλγόριθμο ανίχνευσης κοινοτήτων Louvain [15] στο προκύπτον γράφημα. Οι επιλεγμένες κοινότητες των πειραμάτων μας έχουν επεκταθεί 37% και 34.9% αντίστοιχα από την άποψη του αριθμού των κόμβων. Επίσης, χρησιμοποιούμε τη διαδικασία εντοπισμού bot για να αξιολογήσουμε τους εξεταζόμενους κόμβους. Τα αποτελέσματά μας δείχνουν ότι 427 κόμβοι της κοινότητας χαρακτηρίζονται ως εξαίρετοι και 462 κόμβοι της δεύτερης κοινότητας έχουν συνολική βαθμολογία πάνω από 0.6. Αυτοί οι λογαριασμοί αφαιρούνται από τα ακόλουθα βήματα της ανάλυσης. Το ποσοστό ανίχνευσης γέφυρας στην περίπτωση των 2 κοινοτήτων είναι περίπου 0.8%.

Στοχεύουμε στους χρήστες ανάλογα με τη συμμετοχή τους σε γέφυρες που υπάρχουν ήδη στο γράφημα ή δημιουργούνται κατά τη διάρκεια της εκστρατείας και/ή κατά τη διάρκεια της εκστρατείας μάρκετινγκ. Συσχετίζουμε την μετρική μας, bridge participation, με αυτή του Klout που χρησιμοποιήσαμε σε πολλές μελέτες που μετράνε την επιρροή του χρήστη σε ένα κοινωνικό δίκτυο. Παρατηρούμε υψηλή γραμμική συσχέτιση Spearman στην περίπτωση των κορυφαίων 500 χρηστών της επιρροής με το συντελεστή συσχέτισης πολύ κοντά στα 0.6 κατά τη διάρκεια της εκστρατείας και ακόμα περισσότερο για τους 1000 κορυφαίους χρήστες παραμένει θετικός λίγος συσχετισμός με την βαθμολογία συσχέτισης πολύ κοντά στα 0.6. Και στις δύο περιπτώσεις, η τιμή m είναι περίπου μηδέν, γεγονός που υποδηλώνει υψηλή στατιστική εμπιστοσύνη στα αποτελέσματα.
Α’.6 Διείσδυση Αυτοματοποιημένων Μηχανών

Οι εταιρείες προωθούν τα προϊόντα και το διαφημιστικό τους περιεχόμενο στο ευρύ κοινό και έτσι η ανάπτυξη αυτοματοποιημένων μηχανών ΑΙ (bots) συμβάλλει στην επίτευξη αυτού του στόχου, ανιχνεύοντας τα ενδιαφέροντα των χρηστών και αναπαράγοντας το διαφημιστικό περιεχόμενο μαζικά. Οι εφαρμογές πουκίλουν από τη μαζική χειραγώγηση κατά τη διάρκεια των πολιτικών γεγονότων και εξελίξεων εώς την χρησιμοποίησή τους για εμπορικούς σκοπούς. Επίσης, η χρήση προηγμένων τεχνικών AI και ο συνδυασμός με αποτελεσματικές στρατηγικές μπορούν να μετατρέψουν μια τέτοια μηχανή σε έναν χρήστη επιρροής. Η πρώτη απόπειρα [88] παρουσίασε ένα κοινωνικό AI bot που δημοσίευε αυτοματοποιημένα τωεετς, ακολούθος τους χρήστες στο Τούττερ και περιέγραφε τον εαυτό του ως δημοσιογράφο της Βραζιλίας και έτσι πέτυχε υψηλό βαθμό επιρροής. Μια πρόσφατη μελέτη [47] διερεύνησε και σύγκρινε διάφορες στρατηγικές που εκθέτουν την ευπάθεια του Twitter σε σχέση με επιθέσεις από τέτοιες μηχανές που μπορούν να επηρεάσουν τόσο το ίδιο το Twitter όσο και υπηρεσίες που χτίστηκαν από το Twitter. Μια μελέτη [18] στο Facebook OSN έδειξε ότι οι λογαριασμοί bot ήταν σε θέση να διεισδύσουν επιτυχώς στο δίκτυο σε μια κλίμακα που εξαρτάται από τις ρυθμίσεις απορρήτου των χρηστών. Επιπλέον, το Facebook εισήγαγε νέες λειτουργίες της εφαρμογής τους στο Facebook Messenger-bots.

Ερευνήσαμε αν τα κοινωνικά bots που δημιουργούν αυτόματες συζήτησες μέσω άμεσων μηνυμάτων, μπορούν να διεισδύσουν πιο αποτελεσματικά σε κοινωνικά δίκτυα. Σε προηγούμενες μελέτες, η λειτουργία της συνομιλίας δεν ήταν ενεργοποιημένη υποθέτοντας ότι αυτό θα μπορούσε να μετρήσει αρνητικά και ακόμη να χρησιμοποιηθεί από τους χρήστες ως κριτήριο να ακολουθήσει ή όχι. Τεθεωρούμε ότι οι διαδραστικές ενέργειες βοηθούν να διεισδύσουμε στο κοινωνικό δίκτυο και να δημιουργήσουμε μια δέσμευση μεταξύ bot και χρηστών. Για να το κάνουμε αυτό αναπτύξαμε ένα σύνολο αυτοματοποιημένων χρηστών στο Twitter που χρησιμοποιούν τεχνολογίες αιχμής για να επικοινωνούν και συγκρίνουν τη διείσδυσή τους με μηχανικές χωρίς αυτή την υποστήριξη. Η μελέτη μας εισάγει bots που είναι σε θέση να εκδηλώσουν μια συζήτηση, να στείλουν ένα σχόλιο και να απαντήσουν στα άμεσα μηνύματα. Η μεθοδολογία μας διερεύνα την επιρροή από την άποψη της μετρικής Klout, των αλληλεπιδράσεων μηχανών και του αριθμού των συνδέσεων. Η συμβολή μας δίνει μια εκλεκτική διάσταση χειρισμού κοινωνικών bots. Πρότον, ερευνούμε το επίπεδο διείσδυσης των κοινωνικών bots με λειτουργίες επικοινωνίας υψηλού επιπέδου και εισάγουμε τα Conversational Social Bots (CSB). Δεύτερον, προτείνουμε αυτό το είδος bots ως εργαλείο μάρκετινγκ για διατήρηση της αφοσίωσης.
σε εμπορικά σήματα. Θεωρούμε ότι η ανάλυση της διείσδυσης των κοινωνικών bots και της χρήσης μηχανών AI μπορεί να εισαχάγει μια νέα κατεύθυνση στο μάρκετινγκ κοινωνικών μέσων.

Η μεθοδολογία μας εισάγει AI bots ως υποστηρικτές περιεχομένου. Εξετάζουμε το επίπεδο διείσδυσης των κοινωνικών bots και την επικοινωνιακή απόδοσή τους από την άποψη της επιρροής, των συνδέσεων και των αλληλεπιδράσεων μηνυμάτων. Η μεθοδολογία μας αρχίζει με τη δημιουργία ενός συνόλου κοινωνικών bots. Εστιάζουμε μόνο στο Twitter λόγω του API που μας δίνει την ευκαιρία να εφαρμόσουμε τα κοινωνικά bots με όλες τις διαθέσιμες ενέργειες. Οι διαμορφώσεις και η διαφοροποίηση στις ενέργειες του Twitter είναι άφθονα λαμβάνοντας υπόψη τον αριθμό των επιλογών. Σύμφωνα με τη μελέτη [47], η διαμόρφωση και η δραστηριότητα των κοινωνικών bots καθορίζουν το επίπεδο της διείσδυσής τους στο Twitter. Ωστόσο, στην ανάλυσή τους δεν διαθέτουν προηγμένες δεξιότητες επικοινωνίας που είναι απαραίτητες για την προώθηση του διαφημιστικού περιεχομένου μέσω συνομιλιών.

Δημιουργήσαμε ένα σύνολο λογαριασμών στο Twitter. Τα bots διαφοροποιήθηκαν στις λειτουργίες τους όπως φαίνεται στο Πίνακα Α’.1. Πρώτον, εξετάσαμε το επίπεδο αναγνώρισης των bots από τους κοινωνικούς λογαριασμούς/χρήστες χρησιμοποιώντας το Botometer [37], το οποίο αποτελεί μια πολύ γνωστή υπηρεσία διαδικτύου για την ανίχνευση bots [44, 37, 121]. Με βάση τη συνολική βαθμολογία, τα δημιουργημένα bots δεν είναι αναγνωρίσιμα.

Συμπεραίνουμε ότι το CSB έχει καλύτερες επιδόσεις όσον αφορά την αναλογία Klout, follow και αλληλεπιδράσεις μηνυμάτων. Στα Σχήματα Α’.2,Α’.3 & Α’.4 παρουσιάζεται το ποσοστό βελτίωσης σε σύγκριση με τη μελέτη [47]. Θεωρούμε ότι υπάρχει μια σημαντική βελτίωση. Η 10η ημέρα αποτελεί την κορυφή της βελτίωσης Klout με 37%. Επίσης, η μεγάλη διαφορά μεταξύ του CSB και του βασικού bot είναι στο τέλος των 30 ημερών φθάνοντας στα 123% καλύτερης απόδοσης στην αναλογία follow. Οι αλληλεπιδράσεις μηνυμάτων για το CSB είναι περίπου τέσσερις φορές περισσότερες από το βασικό bot.
Σχήμα Α.2: Βελτίωση % Klout μεταξύ CSB και Baseline Bot

Σχήμα Α.3: Βελτίωση % Follow ratio μεταξύ CSB και Baseline Bot

Α’.7 Εξατομικευμένη διαφήμιση

Στο ψηφιακό μάρκετινγκ, είναι δύσκολο να καλυφθούν οι ανάγκες όλων των χρηστών όταν ο αριθμός των σελίδων και των κατηγοριών σε έναν δυναμικό ιστότοπο αυξάνεται. Πολλοί ιστότοποι περιέχουν εκατοντάδες ή χιλιάδες διαφορετικές κατηγορίες, δημιουργώντας δυσκολίες στην εύρεση της σελίδας που θέλει ο χρήστης, υποβαθμίζοντας την ποιότητα του ιστότοπου. Ταυτόχρονα, η εμπειρία περιήγησης σχετίζεται με τις θετικές ή αρνητικές αντιδράσεις στις διαφημίσεις και ως εκ τούτου η ανακατασκευή ενός ιστότοπου αποτελεί σημαντικό παράγοντα. Ο στόχος της ανακατασκευής ενός ιστότοπου δεν είναι μόνο η πρόσβαση σε δημοφιλείς σελίδες με λιγότερα βήματα, ξεκινώντας
Σχήμα Α’.4: Βελτίωση % Message Interactions μεταξύ CSB και Baseline Bot

ατό την αρχική σελίδα, αλλά και η κάλυψη των αναγκών των χρηστών για νέα προϊόντα. Μια καλά αναφερθείσα μεθοδολογία είναι η χρήση πρόσθετων συνδέσμων, δηλαδή hotlinks [41, 17, 97, 98], που συνδέουν τις δημοφιλείς ιστοσελίδες με τους απογόνους τους, μειώνοντας την απόσταση από την αρχική σελίδα.

Ερευνήσαμε το θέμα της εξατομικευμένης διαφήμισης μέσα σε έναν δυναμικό ιστότοπο. Θεωρούμε ότι η ικανοποίηση των χρηστών κατά την περιήγηση σχετίζεται με την πιθανότητα του καταναλωτή να αγοράσει ένα προϊόν και να υποστηρίξει μια εκστρατεία μάρκετινγκ. Η μεθοδολογία μας εστιάζει στο διαφημιστικό περιεχόμενο και δημιουργεί συνδέσμους (hotlinks). Τα βασικά σημεία της συμβολής μας μπορούν να συνοψιστούν στις ακόλουθες προτάσεις. Χρησιμοποιούμε την έννοια της ανάθεσης hotlink για την ανακατασκευή του ιστότοπου. Ένα κοινό χαρακτηριστικό των προηγουμένων μελετών είναι ότι ασχολούνται με τη δημοτικότητα των σελίδων χωρίς να λαμβάνονται υπόψη οι πληροφορίες από τα κοινωνικά μέσα. Σημειώνεται ότι οι προηγούμενες εργασίες ασχολούνται με τη δημοτικότητα μιας ιστοσελίδας μόνο με τον αριθμό των clicks που οι σελίδες λαμβάνουν από τους χρήστες κατά την περιήγηση. Επιπλέον, με βάση τις γνώσεις μας, αυτή η μελέτη είναι η πρώτη που εισάγει ως κριτήριο στην απόφαση την πολυπλοκότητα των ιστότοπων. Σημειώνεται ότι η ανάθεση εξατομικευμένης διαφήμισης μεταξύ μιας ιστοσελίδας και της ιδιαίτερης μετρική Pagerank που λαμβάνει υπόψη την σημασιολογική ομοιότητα και πολυπλοκότητα των σελίδων στην κατάταξή τους.
α’.7. Εξατομικευμένη διαφήμιση

μένου, δηλαδή διαφημιστικά banners, βίντεο, συνδέσμους σε έναν δυναμικό ιστότοπο. Η διαδικασία περιλαμβάνει μια ολιστική ανακατασκευή δικτυακού τόπου μέσω της ανάθεσης επιπρόσθετων συνδέσμων hotlink όπως παρέχουν οι μελέτες [83, 84]. Παρουσιάζουμε έναν αλγόριθμο αντιστοίχισης hotlink προκειμένου να βελτιώσουμε την εμπειρία περιήγησης με βάση τη μείωση της απόστασης μεταξύ των δημοφιλών σελίδων και της σελίδας προορισμού. Υποστηρίζουμε αυτή τη διαδικασία που συνδέει τις σελίδες με την ομοιότητα στο περιεχόμενο μεταξύ των σελίδων και την μείωση της συναλλαγής πολυπλοκότητα του ιστότοπου. Επίσης, εκμεταλλευόμαστε τις διαθέσιμες πληροφορίες από τα κοινωνικά μέσα, προκειμένου να εξατομικεύσουμε τη δημιουργία hotlinks με βάση τα ενδιαφέροντα των χρηστών.

Στο Σχήμα Α’.5 αναλύονται λεπτομερώς οι λειτουργικές μονάδες του μοντέλου μας και η αρχιτεκτονική του συστήματος.

Μια σύνοψη των ανεξάρτητων εργασιών που περιλαμβάνει η μεθοδολογία μας δίνεται παρακάτω:

- **Δημιουργία γράφηματος ιστότοπου.** Δημιουργούμε ένα κατευθυνόμενο γράφημα $G(V,E)$ για να μοντελοποιήσουμε του ιστότοπο όπου οι κόμβοι και οι άκμες είναι σελίδες και συνδέσμοι προς τον ιστότοπο αντίστοιχα. Δηλώνουμε ένα χαρακτηριστικό δημοτικότητας σε κάθε κόμβο και 2 ανεξάρτητα βάρη, δηλαδή μια ομοιότητα περιεχομένου και ένα βάρος αύξησης φορτίου σελίδας. Επίσης, με βάση τις σημασιολογικές πληροφορίες, εξάγεται η κατηγορία κάθε ιστοσελίδας, δημιουργώντας μια λίστα κατηγοριών για ολόκληρο τον ιστότοπο.

- **Εξαγωγή των προτιμήσεων του χρήστη.** Αυτή η μονάδα είναι αφιερωμένη στην παραγωγή μιας λίστας κατηγοριών που προέρχονται από τα κείμενα των χρηστών στα κοινωνικά μέσα. Αυτή η εργασία λειτουργεί μόνο όταν ο χρήστης είναι ενεργός στο κοινωνικό δίκτυο.

- **Δημιουργία Hotlinks.** Ο αλγόριθμος δημιουργεί πρόσθετους συνδέσμους για έναν κόμβο σε έναν από τους απογόνους του, για να μειώσει την απόσταση από την αρχική σελίδα στην σελίδα στόχο βελτιώνοντας την περιήγησης.

- **Ανακατασκευή Ιστοσελίδων & Τοποθέτηση Διαφημίσεων.** Αυτή η μονάδα είναι υπεύθυνη για την ανακατασκευή ιστότοπου, λαμβάνοντας υπόψη τη δημοτικότητα των σελίδων, την ομοιότητα περιεχομένου, την απόδοση φορτίου και τα προσωπικά ενδιαφέροντα.
Η καινοτόμος μεθοδολογία μας βελτιώνει την περιήγηση των χρηστών αφού σε λιγότερα βήματα δέχονται τις πληροφορίες που συνδέονται με το διαφημιστικό περιεχόμενο. Προσαρμόζουμε μια ολιστική διαδικασία ανακατασκευής ιστοσελίδων για να τοποθετήσουμε το διαφημιστικό περιεχόμενο με αποτελεσματικό τρόπο και να βελτιώσουμε την απόδοσή της εκστρατείας μάρκετινγκ. Τονίζουμε ότι η χρήση των κοινωνικών μέσων συνεισφέρει στον επανυπολογισμό της δημοτικότητας κάθε ιστοσελίδας. Σημειώνεται ότι οι περισσότερες από τις μελέτες στο στην εκχώρηση πρόσθετων συνδέσμων επικεντρώνουν στον αριθμό των προσβάσεων κατά την περιήγηση από τους χρήστες ως καθοριστικό παράγοντα για τη μέτρηση της δημοτικότητας μιας ιστοσελίδας. Επίσης πολλές φορές οι χρήστες αισθάνονται δυσαρεστημένοι με την απώλεια χρόνου και συνεπώς η μείωση των βήματων και της ταχύτητας ψήφωσης της σελίδας αποτελεί σημαντικό παράγοντας αισθήματος των χρηστών. Με βάση τις γνώσεις μας, αυτή η μελέτη είναι η πρώτη που συνδυάζει την πολυπλοκότητα των δικτυακών τόπων και εισάγει αυτό το κριτήριο στην απόφαση. Επιπλέον, προτείνουμε μια ελαφρώς διαφορετική μετρική Pagerank που λαμβάνει υπόψη την σημασιολογική ομοιότητα και πολυπλοκότητα των σελίδων στην κατάταξη. Σύμφωνα με την πειραματική μας διαδικασία, η προτεινόμενη μεθοδολογία βελτιώνει την εμπειρία περιήγησης από την ύπαρξη της απόστασης και μειώνει περίπου τα 11% τα βήματα για την επίτευξη του στόχου της ιστοσελίδας. Επίσης, στα πειράματα μας μειώνουμε την απώλεια χρόνου και μνήμης σε περίπου 17.5% και 20% αντίστοιχα κατά την περιήγηση.
Διάχυση πληροφορίας στα Κοινωνικά Μέσα

Α’.8 Διάχυση πληροφορίας στα Κοινωνικά Μέσα

Σε μια καμπάνια μάρκετινγκ, μας ενδιαφέρει να εντοπίσουμε τους αρχικούς χρήστες που θα μεταδώσουν το διαφημιστικό μήνυμα ώστε να επιτευχθεί η μέγιστη διάχυση. Αυτό το στάδιο ονομάζεται διάδοση στα Κοινωνικά Μέσα στο δίκτυο, η επικοινωνία των πελατών έχει ως αποτέλεσμα τη διάδοση πληροφοριών. Προτείνουμε έναν καινοτόμο τρόπο ανακάλυψης αυτών των αρχικών χρηστών, χρησιμοποιώντας πληροφορίες από το ιστορικό τους σχετικά με τη διάχυση πληροφορίας, Στην τρέχουσα βιβλιογραφία, η συμπεριφορά των καταναλωτών και οι καταναλωτικές συνήθειες έχουν προταθεί για την αναγνώριση αυτών των χρηστών. Στην προσέγγισή μας, βασίζομαι ενεργοποιημένοι κόμβοι καλούμε την πολιτική μας από ένα διαφορετικό σύνολο δεδομένων πραγματικής συμμετοχής στο κοινωνικό δίκτυο.

Η διαδικασία διάχυσης της πληροφορίας που προσομοιώνει την πραγματική εξάπλωση, έχει λάβει μεγάλη προσοχή στη βιβλιογραφία χρησιμοποιώντας κυρίως πιθανοτικά μοντέλα. Έχει συσχετιστεί με μόλυνση από ιούς και έχει αναπτυχθεί πλήθος εργαλείων και μοντέλων που προσπαθούν να εξηγήσουν την εξάπλωση των πληροφοριών [5, 128]. Η δυνατότητα να αξιοποιηθεί αυτόματα μεγάλος όγκος πληροφοριών διαφορετικού τύπου από τα κοινωνικά μέσα, δίνει μια νέα διάσταση στην επεξεργασία της επικοινωνικής συμπεριφοράς των χρηστών. Υιοθετούμε μια νέα προσέγγιση στην διάχυση πληροφοριών και προτείνουμε την αποτελεσματική χρήση ενός τοιχίου συνόλου δεδομένων, δηλαδή την πραγματική δομή του δικτύου, των μετρήσεων συμπεριφοράς του χρήστη, του χώρου και των αναφερόμενων για ένα συγκεκριμένο θέμα αναπτύσσοντας μια πολιτική εφαρμογή της διάχυσης.

Η κύρια συμβολή μας είναι η εκμετάλλευση του μοντέλου διάχυσης, στην επιλογή των κόμβων που επιτυγχάνουν την καλύτερη κάλυψη του γραφήματος. Η κάλυψη χαρακτηρίζει τον αριθμό των κόμβων που έλαβαν τη δοσμένη πληροφορία. Αυτοί οι κόμβοι καλούνται ενεργοποιημένοι κόμβοι. Προτείνουμε τον αλγόριθμο βασισμένο σε μοτίβα διάδοσης πληροφοριών. Αυτό μας επιτρέπει να προσδιορίζουμε τα βασικά χαρακτηριστικά διάχυσης των χρηστών με πραγματικό τρόπο. Για να επικυρώσουμε τα αποτελέσματα μας, υποδηλώνει την προσέγγιση μας με τις κύριες υπάρχουσες λύσεις. Τα αποτελέσματα μας δείχνουν ότι μια εκστρατεία μάρκετινγκ που παρέχει κίνητρα στον χώρο των χρηστών του προτείνουμε μπορεί να επιτύχει παρόμοια αποτελέσματα με ένα πιθανοτικό μοντέλο διάδοσης. Επιπλέον, για να επικυρώσουμε περαιτέρω την προσέγγισή μας, χρησιμοποιούμε μια μεθοδολογία για τη μέτρηση της πραγματικής διάδοσης του δικτύου.
Θεωρούμε ότι ένα από τα σημαντικότερα καθήκοντα στις εκστρατείες μάρκετινγκ είναι η διάδοση συγκεκριμένων διαφημίσεων ή προσφορών σε μεγάλο αριθμό θεατών προκειμένου να μεγιστοποιηθεί το ενδεχόμενο ένα μέρος ή το σύνολο αυτών να είναι μελλοντικοί πελάτες του προϊόντος. Λόγω του προϋπολογισμού και των χρονικών περιορισμών, αυτή η διάδοση δεν είναι εφικτή στέλνοντας ένα μήνυμα σε όλους τους χρήστες των κοινωνικών μέσων, αλλά επιλέγοντας τους κόμβους που μπορούν να διαχέουν αποτελεσματικά το μήνυμα στις συνδέσεις τους. Με βάση αυτή την έννοια, η μεθοδολογία μας εισάγει έναν αλγόριθμο διάχυσης βασισμένο σε πρότυπα, ο οποίος προβλέπει το μοτίβο διάχυσης κάθε χρήστη και εισάγει μια διαδικασία διάδοσης παρόμοια με τις εργασίες [62, 63].

Σχήμα Α’.6: Αρχιτεκτονική συστήματος διάχυσης βασισμένο σε πρότυπα

Μια σύνοψη των ανεξάρτητων εργασιών που περιλαμβάνει η μεθοδολογία μας δίνεται στο Σχήμα Α’.6 και στα παρακάτω σημεία:

- Ανίχνευση δεδομένων. Συγκέντρωση των δεδομένων από το διαδικτυακό κοινωνικό δίκτυο Twitter που είναι απαραίτητη για τις εργασίες, συμπεριλαμβανομένων των κειμένων, της λίστας των φίλων και των μετρικών κοινωνικής συμπεριφοράς.

- Προφίλ χρήστη. Οι κατάλληλες πληροφορίες για κάθε χρήστη συγκεντρώνονται και ένα διάνυσμα χαρακτηριστικών δημιουργείται για κάθε μοναδικό χρήστη του συνόλου δεδομένων.
• Πρόβλεψη του μοτίβου επαναδημοσίευσης. Ένα μοντέλο πρόβλεψης εκπαιδεύεται για να κατηγοριοποιήσει τη διάχυση. Κάθε χρήστης κατηγοριοποιείται σε μία συγκεκριμένη δομή διάχυσης.

• Διάχυση βασισμένη σε πρότυπα. Επιλογή των καλύτερων κόμβων για τη διάδοση μηνυμάτων από άποψη χρόνου και κάλυψης του κοινωνικού γραφήματος.

Με βάση τις τρέχουσες θεωρίες διάχυσης, χρησιμοποιώντας πιθανοτικά μοντέλα, οι αλγόριθμοι μπορούν να εκτιμήσουν τη διάχυση των πληροφοριών. Αλλά όταν πρόκειται να επηρεάσει τη διάχυση δεν υπάρχουν πολλές πληροφορίες. Η μεθοδολογία μας προτείνει έναν τρόπο επίτευξης ενός ποσοστού διάχυσης όπως αυτός που προβλέπεται από το IC.

Ανακαλύπτουμε τους κόμβους του γραφήματος που μπορούν να μειώσουν τη διάχυση δίνοντας ένα συγκεκριμένο τύπο μηνυμάτων. Τα πραγματικά δεδομένα χρησιμοποιούν μετρήσεις γλωσσολογίας και κοινωνικής συμπεριφοράς σε ένα χρονικό παράθυρο και έτσι οι τρέχουσες αλλαγές στο μοτίβο διάχυσης ίσως δεν είναι ορατές. Ωστόσο, το διάνυσμα χαρακτηριστικών μπορεί να τροφοδοτείται με τα τρέχοντα κείμενα. Το διάνυσμα χαρακτηριστικών έχει δημιουργηθεί με βάση τη γλωσσολογία και συναισθηματικά χαρακτηριστικά που προέρχονται από το λογισμικό LIWC [95] ώστε να βρίσκονται κοντά σε έναν πολυδιάστατο χώρο. Η πρόβλεψη του μοντέλου διάχυσης για κάθε χρήστη χρησιμοποιείται στον αλγόριθμο μας (PBD). Δεδομένης μας συγκεκριμένης δομής δικτύου, μπορούμε να συμπεράνουμε ποιο είναι το πρότυπο διάχυσης που μπορεί να επιτύχει την υψηλότερη διάχυση. Με δεδομένες αυτές τις πληροφορίες, μια καμπάνια αρχικών καμπάνια μπορεί να ξεκινήσει την εντοπισμό κ αρχικών προτύπων L3 και στη συνέχεια να δώσει κίνητρα σε όλες τις κοινωνικές δομές στο δίκτυο. Ως εκ τούτου, οι άτομα που διαχειρίζονται τις εκστρατείες θα πρέπει να εστιάσουν τις προσπάθειες τους για να ενεργοποιήσουν τους χρήστες με προηγούμενη διάχυση προτύπου L3.
Από την άλλη πλευρά, όταν χρησιμοποιούνται όλα τα πρότυπα ($L_2 - L_3 - L_4$), το $PBD$ μπορεί να επιτύχει διάχυση παρόμοια με αυτή του $LT$. Είναι ενδιαφέρον να σημειώσουμε ότι ο αλγόριθμος $IC$ μπορεί να οδηγήσει σε διαφορετικό ποσοστό κάλυψης γραφήματος φτάνοντας ακόμα και την πλήρη κάλυψη με βάση την τιμή $p$. Όπως ήδη αναφέρθηκε, εμείς πειραματίζουμε με $p$ όπως παρουσιάζεται στο [62] και όπως επιβεβαιώνεται από την πραγματική μεθοδολογία διάχυσης. Ακόμα, το γεγονός αυτό ενισχύει το επιχείρημα ότι η προσέγγιση μας που δεν βασίζεται σε κάποια ρύθμιση παραμέτρων αλλά στη συμπεριφορά των προηγούμενων χρηστών που είναι μια ρεαλιστική προσέγγιση για τη διάδοση της διάδοσης πληροφοριών μέσα σε ένα δίκτυο. Τα πιθανοτικά μοντέλα $IC$ και $LT$ διερευνούν τη διάχυση με βάση τις πιθανότητες ενεργοποίησης στα άκρα. Στα πειράματα μας χρησιμοποιήσαμε την τιμή $p = 0,1$ για το $IC$ όπως χρησιμοποιείται στο [62]. Στην προσπάθειά μας να επαληθεύσουμε ότι η $IC$ επιτυγχάνει μια ρεαλιστική διάχυση, αναπτύξαμε μια μεθοδολογία μέτρησης της πραγματικής διάδοσης. Τα αποτελέσματα μας δείχνουν ότι για το σύνολο δεδομένων μας η πραγματική διάχυση του μηνύματος έγινε από το 19% των χρηστών, το οποίο είναι το ίδιο αποτέλεσμα που παράγεται από το $IC$ και από το $PBD/L3$. Αυτό ενισχύει την εγκυρότητα της μεθοδολογίας μας.

Ως μελλοντική εργασία, εξετάζουμε το ενδεχόμενο να διερευνήσουμε τη συνδυασμό $PBD$ με πιθανοτικά μοντέλα και να πειραματιστούμε στη δημιουργία διαφορετικών ηευριστικών μοντέλων με βάση το $LT$ ή το $IC$. Επιπλέον, η τεχνική μας θα μπορούσε να επεκτείνεται σε πολύγλωσσο μοντέλο χρησιμοποιώντας έναν κοινοτικό αλγόριθμο ανίχνευσης καλύτερων κ-κόμβων από τους εκπροσώπους κάθε κοινότητας. Επιπλέον, η τεχνική μας θα μπορούσε να επεκτείνεται σε πολύγλωσσο μοντέλο σε διάφορες κοινότητες του κοινωνικού γραφήματος και συγχωνεύοντας τα αποτελέσματα σε ολόκληρο το γράφημα για να επιτύχουμε καλύτερες επιδόσεις.

Α'9 Μελλοντική Εργασία

Ως μελλοντική εργασία, ενδιαφέρομαστε να αναπτύξουμε τις παραπάνω μεθοδολογίες σε επισκευασμένους κοινωνικά δίκτυα όπως το Facebook, Linkedin και το Instagram και να αναγνωρίσουμε τους τρόπους παράγοντες που επηρεάζουν τα αποτελέσματα των ολοκλήρως μας.

Επιπλέον, η προσαρμογή της προτεινόμενης μεθοδολογίας σε διάφορους τύπους εκστρατειών μάρκετινγκ, π.χ. εκλογές, προσφορές προϊόντων, προσφορές εργασίας.

Στο επίπεδο της εξατομίκευσης η μελέτη μας μπορεί να εξάγει τις προτιμήσεις των χρηστών σε μία μόνο γλώσσα, όμως το σχέδιο μας είναι να επεκτείνουμε την τεχνική μας σε πολύγλωσσο μοντέλο.
Όσον αφορά την πρόβλεψη συνδέσης θεωρούμε ότι ο προσδιορισμός της θέσης του
χρήστη θα μπορούσε να βελτιώσει την απόδοση.

Εξετάζουμε το ενδεχόμενο να διερευνήσουμε τον συνδυασμό του αλγορίθμου δι-
άγμασης με πιθανοτικά μοντέλα. Επιπλέον, η τεχνική μας θα μπορούσε να επεκταθεί
χρησιμοποιώντας έναν αλγόριθμο εύρεσης κοινοτήτων έτσι ώστε να επιλέγονται οι εκ-
προσώποι κάθε κοινότητας.

Τέλος, στα τρέχοντα ενδιαφέροντά μας είναι η χρήση αποθήκης τεχνολογιών cloud
computing και ο πειραματισμός των υπολογιστικών ζητημάτων σε μεγάλου όγκου δε-
δομένα.
Bibliography


[8] Eytan Bakshy, Brian Karrer, and Lada A Adamic. Social influence and the
diffusion of user-created content. In Proceedings of the 10th ACM conference

[9] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of
social networks in information diffusion. In Proceedings of the 21st international

Markov constraints for generating lyrics with style. In Proceedings of the 20th

Detecting spammers on twitter. In Collaboration, electronic messaging, anti-

[12] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 u.s. presiden-
tial election online discussion. First Monday, November 2016.

maximization in social networks. Internet and Network Economics, pages 306–

[14] Parantapa Bhattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi
Ghosh, and Krishna P Gummadi. Inferring user interests in the twitter social
network. In Proceedings of the 8th ACM Conference on Recommender systems,

Lefebvre. Fast unfolding of communities in large networks. Journal of sta-

[16] Allan Borodin, Yuval Filmus, and Joel Oren. Threshold models for competitive
influence in social networks. In WINE, volume 6484, pages 539–550. Springer,
2010.

[17] Prosenjit Bose, Jurek Czyzowicz, Leszek Gasieniec, Evangelos Kranakis, Danny
Krizanc, Andrzej Pelc, and Miguel Vargas Martin. Strategies for hotlink as-
signments. In International Symposium on Algorithms and Computation, pages


[113] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.


