

# Information Diffusion and Summarization in Social Networks

Nagendra Kumar

A Thesis Submitted to  
Indian Institute of Technology Hyderabad  
In Partial Fulfillment of the Requirements for  
The Degree of Doctor of Philosophy



भारतीय प्रौद्योगिकी संस्थान हैदराबाद  
Indian Institute of Technology Hyderabad

Department of Computer Science and Engineering

August 2019

## Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

Nagendra Kumar

(Signature)

Nagendra Kumar


(Name)

CS14RESCH11005

(Roll No.)

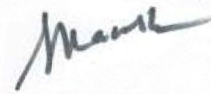
## Approval Sheet

This thesis entitled by **Information Diffusion and Summarization in Social Networks** by **Nagendra Kumar** is approved for the degree of Doctor of Philosophy from IIT Hyderabad.

  
2/9/17

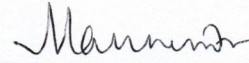
Professor P Sreenivasa Kumar, Dept. of CSE IITM

Examiner 1



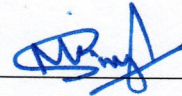
Professor Saroj Kaushik, Dept. of CSE IITD

Examiner 2



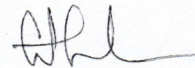
Dr. Maunendra Sankar Desarkar, Dept. of CSE IITH

Internal Examiner



Dr. Manish Singh, Dept. of CSE IITH

Adviser/Guide



Dr. Sumohana S. Channappayya, Dept. of EE IITH

Chairman

## Acknowledgements

I would like to take the opportunity to thank everyone who directly and indirectly helped me in realizing the thesis. First of all, I would like to thank the almighty God for giving me strength, grace, and wisdom to pursue my P.h.D. I would like to express my sincere gratitude to my adviser, Dr. Manish Singh for his unconditional support. He is the one who trained me to carry out impactful research. This thesis was not possible without his support. He did not support me only in technical discussions but also helped me to become a better person. Further, I would like to thank my parents, brother, and uncle who ceaselessly supported me throughout my life. I would like to especially thank my younger brother who was always there with me in my tough and good time of life. He encouraged me to do Ph.D. and motivated all the time during my P.h.D. I would also like to thank my doctoral committee members Dr. Maunendra Sankar Desarkar, Dr. Manohar Kaul and Dr. Sumohana S. Channappayya who encouraged me and provided astute suggestions which widened my thesis from multiple perspectives. Last but not least, I would also like to thank all the co-authors and fellow labmates especially Anand Konjengbam who helped in technical discussions and writing research papers.

**Nagendra Kumar**

# Abstract

Social networks are web-based services that allow users to connect and share information. Due to the huge size of social network graph and the plethora of generated content, it is difficult to diffuse and summarize the social media content. This thesis thus addresses the problems of information diffusion and information summarization in social networks.

Information diffusion is a process by which information about new opinions, behaviors, conventions, practices, and technologies flow from person-to-person through a social network. Studies on information diffusion primarily focus on how information diffuses in networks and how to enhance information diffusion. Our aim is to enhance the information diffusion in social networks. Many factors affect information diffusion, such as network connectivity, location, posting timestamp, post content, etc. In this thesis, we analyze the effect of three of the most important factors of information diffusion, namely network connectivity, posting time and post content. We first study the network factor to enhance the information diffusion, and later analyze how time and content factors can diffuse the information to a large number of users.

Network connectivity of a user determines his ability to disseminate information. A well-connected authoritative user can disseminate information to a more wider audience compared to an ordinary user. We present a novel algorithm to find topic-sensitive authorities in social networks. We use the topic-specific authoritative position of the users to promote a given topic through word-of-mouth (WoM) marketing. Next, the lifetime of social media content is very short, which is typically a few hours. If post content is posted at the time when the targeted audience are not online or are not interested in interacting with the content, the content will not receive high audience reaction. We look at the problem of finding the best posting time(s) to get high information diffusion. Further, the type of social media content determines the amount of audience interaction, it gets in social media. Users react differently

to different types of content. If a post is related to a topic that is more arousing or debatable, then it tends to get more comments. We propose a novel method to identify whether a post has high arousal content or not. Furthermore, the sentiment of post content is also an important factor to garner users' attention in social media. Same information conveyed with different sentiments receives a different amount of audience reactions. We understand to what extent the sentiment policies employed in social media have been successful to catch users' attention.

Finally, we study the problem of information summarization in social networks. Social media services generate a huge volume of data every day, which is difficult to search or comprehend. Information summarization is a process of creating a concise readable summary of this huge volume of unstructured information. We present a novel method to summarize unstructured social media text by generating topics similar to manually created topics. We also show a comprehensive topical summary by grouping semantically related topics.

**Keywords:** Social network analysis, Information diffusion, Information summarization, Text mining, Influential users, Data characterization

# Contents

Declaration . . . . .	ii
Approval Sheet . . . . .	iii
Acknowledgements . . . . .	iv
Abstract . . . . .	v
List of Abbreviations . . . . .	xv
<b>1 Introduction</b>	<b>1</b>
1.1 Information Diffusion . . . . .	2
1.1.1 Determinants of Information Diffusion . . . . .	2
1.2 Information Summarization . . . . .	7
<b>2 Literature Review</b>	<b>10</b>
2.1 Information Diffusion . . . . .	10
2.1.1 Influence of Users in OSNs . . . . .	11
2.1.2 Right Time to Post . . . . .	12
2.1.3 Popularity Prediction of Posts . . . . .	13
2.2 Information Summarization . . . . .	16
2.2.1 Key-topics of Interests . . . . .	16
<b>3 Information Diffusion through Topic-sensitive WoM Marketing</b>	<b>18</b>
3.1 Introduction . . . . .	18
3.2 Problem Definition . . . . .	21

3.3	Analysis of Online Social Groups . . . . .	23
3.4	Social Interaction Graph . . . . .	24
3.4.1	Measuring Topical Relevance . . . . .	25
3.5	Finding Influential Users in OSG . . . . .	29
3.6	Reinforced Marketing . . . . .	29
3.7	Evaluations . . . . .	30
3.7.1	Experimental Setup . . . . .	30
3.7.2	Evaluation Metrics . . . . .	31
3.7.3	Effectiveness of Algorithms . . . . .	32
3.7.4	Precision Analysis . . . . .	34
3.7.5	Marketing Across Topics . . . . .	35
3.7.6	Empirical Evaluation . . . . .	37
3.7.7	Temporal Dynamics . . . . .	38
3.8	Conclusion . . . . .	40
<b>4</b>	<b>Information Diffusion using the Best Time to Post</b>	<b>41</b>
4.1	Introduction . . . . .	41
4.2	Problem Formulation . . . . .	44
4.2.1	Problem Definition . . . . .	44
4.2.2	Dataset . . . . .	46
4.3	Audience Reaction Analysis . . . . .	47
4.3.1	Post to Reaction Time Analysis . . . . .	47
4.3.2	Audience Reaction Behavior Analysis . . . . .	48
4.4	Categorization of Pages . . . . .	49
4.4.1	Reaction Determining Features . . . . .	49
4.4.2	Feature Pre-processing . . . . .	50
4.4.3	Categorization . . . . .	51
4.5	Schedule Derivation . . . . .	52



4.5.1	Aggregated Schedules . . . . .	53
4.5.2	Categorized Schedules . . . . .	55
4.5.3	Weighted Categorized Schedules . . . . .	56
4.6	Evaluations . . . . .	57
4.6.1	Evaluation Metrics . . . . .	57
4.6.2	Effect of Schedule . . . . .	60
4.6.3	Effectiveness of Categorization . . . . .	61
4.6.4	Trend Analysis . . . . .	62
4.6.5	Audience Engagement with Contents . . . . .	66
4.7	Conclusion . . . . .	67
<b>5</b>	<b>Information Diffusion by Posting High Arousal Content</b>	<b>68</b>
5.1	Methodology . . . . .	70
5.1.1	Labeling High and Low Arousal Posts . . . . .	70
5.1.2	Generate the Candidate Features . . . . .	73
5.1.3	Feature Selection . . . . .	75
5.1.4	Arousal based Post Classification . . . . .	76
5.1.5	Determining the Topics of High Arousal . . . . .	77
5.2	Evaluations . . . . .	78
5.2.1	Experimental Setup . . . . .	78
5.2.2	Effectiveness of Methods . . . . .	79
5.2.3	Analyzing the Topics of High Arousal . . . . .	81
5.3	Conclusion . . . . .	82
<b>6</b>	<b>Information Diffusion of News in OSNs using Sentiment Dynamics</b>	<b>84</b>
6.1	Introduction . . . . .	84
6.2	Methodology . . . . .	88
6.2.1	News Posts Collection . . . . .	88

6.2.2	Sentiment Polarity Identification . . . . .	90
6.2.3	News Posts Categorization . . . . .	92
6.3	Analysis of News Posts Polarity . . . . .	94
6.3.1	News Posts Polarity across Categories . . . . .	96
6.3.2	Big Headlines Versus Niche News . . . . .	98
6.3.3	Polarity of Same News Events across Channels . . . . .	100
6.4	Popularity Versus Polarity . . . . .	102
6.5	User Opinion Analysis . . . . .	104
6.6	Temporal Analysis . . . . .	107
6.7	Conclusion . . . . .	109
<b>7</b>	<b>Information Summarization by Generating Topics of Interest</b>	<b>111</b>
7.1	Introduction . . . . .	111
7.2	Methodology . . . . .	114
7.2.1	Data Categorization . . . . .	115
7.2.2	Candidate Topic Generation . . . . .	116
7.2.3	Topic Pruning . . . . .	117
7.2.4	Topic Refining . . . . .	121
7.2.5	Topic Grouping . . . . .	124
7.3	Experimental Evaluations . . . . .	124
7.3.1	Experimental Setup . . . . .	124
7.3.2	Performance Evaluation . . . . .	125
7.3.3	Empirical Analysis . . . . .	129
7.4	Conclusion . . . . .	132
<b>8</b>	<b>Conclusion and Future Work</b>	<b>134</b>
8.1	Summary of the Thesis . . . . .	134
8.1.1	Information Diffusion using WoM Marketing . . . . .	134

8.1.2	Information Diffusion using the Best Time to Post . . . . .	135
8.1.3	Information Diffusion by Posting High Arousal Content . . . .	136
8.1.4	Sentiment Dynamics in Social Media Channels . . . . .	136
8.1.5	Information Summarization by Generating Key-topics . . . . .	136
8.2	Future Work . . . . .	137

<b>References</b>		<b>138</b>
-------------------	--	------------

# List of Figures

1.1	Network structure of a online social group . . . . .	3
1.2	Average reactions per post in different time intervals . . . . .	5
3.1	Bow-tie Structure of OSGs . . . . .	23
3.2	Degree distribution in groups . . . . .	24
3.3	Boosted relevance . . . . .	27
3.4	Correlation of authority measure algorithms with votes . . . . .	32
3.5	Correlation of authority measure algorithms with topical votes . . . . .	33
3.6	Correlation of top users across variety of topics . . . . .	36
3.7	Posting behavior of top users . . . . .	39
3.8	Reaction behavior of top users . . . . .	39
4.1	Audience reaction behavior . . . . .	48
4.2	Reaction Gain . . . . .	60
4.3	Audience reaction pattern on daily basis . . . . .	63
4.4	Audience reaction pattern on weekly basis . . . . .	64
4.5	Audience reaction pattern on monthly basis . . . . .	65
5.1	Overview of the arousal prediction . . . . .	70
6.1	Distribution of news posts across categories . . . . .	94
6.2	Polarity of news posts generated by pages . . . . .	95
6.3	Category-wise post distribution of Fox News . . . . .	97

6.4	Category-wise post distribution of The Economist . . . . .	97
6.5	Category-wise post distribution of NPR . . . . .	97
6.6	Big headlines . . . . .	98
6.7	Niche news . . . . .	99
6.8	Percentage of positive news generated for a news event . . . . .	100
6.9	Percentage of negative news generated for a news event . . . . .	100
6.10	Likes on posts with different sentiments . . . . .	102
6.11	Comments on posts with different sentiments . . . . .	103
6.12	Shares on posts with different sentiments . . . . .	103
6.13	Avg. comment polarity vs. post polarity on CNN . . . . .	104
6.14	Avg. comment polarity vs. post polarity on Fox News . . . . .	104
6.15	Avg. comment polarity vs. post polarity on The Economist . . . . .	105
6.16	Avg. comment polarity vs. post polarity on The New York Times . . . . .	105
6.17	Avg. comment polarity vs. post polarity on NPR . . . . .	105
6.18	Temporal sentiment pattern of The Economist . . . . .	107
6.19	Temporal sentiment pattern of NPR . . . . .	107
6.20	Temporal sentiment pattern of Fox News . . . . .	108
6.21	Temporal sentiment pattern of CNN . . . . .	108
7.1	Topics of interest from VLDB 2017 conference website . . . . .	112
7.2	System architecture to find topics of interest . . . . .	114

# List of Tables

3.1	MAP and NDCG of authority measure algorithms . . . . .	34
3.2	Correlation in top users ranking for popular topics . . . . .	35
4.1	Notations . . . . .	53
4.2	Correlation across the categories . . . . .	61
4.3	Correlation within the category . . . . .	61
4.4	Posts and reactions of different types of contents . . . . .	66
5.1	Performance evaluation of arousal prediction . . . . .	80
5.2	Topics of high arousal from different categories . . . . .	81
6.1	Polarity of a news post generated by two different types of channels .	85
6.2	Dataset statistics . . . . .	89
6.3	Sentiment polarity of sample posts . . . . .	91
6.4	Correlation between post sentiment and comment sentiment . . . . .	106
7.1	MAP of different topic modeling algorithms . . . . .	126
7.2	Similarity of topics with manually listed topics . . . . .	128
7.3	Key-topics of Machine Learning area . . . . .	130
7.4	Topic evolution in Databases over the years . . . . .	131
7.5	Topic Groups of Machine Learning research . . . . .	132

## List of Abbreviations

AFP	Aggregated frequent posting schedule
AFR	Aggregated frequent reaction schedule
API	Application program interface
Avg.	Average
B2P	Business-to-people
BoW	Bag-of-words
CFP	Categorized frequent posting schedule
CFR	Categorized frequent reaction schedule
Entertain	Entertainment
GMT	Greenwich mean time
IC	Independent cascade
IDF	Inverse document frequency
KNN	K-Nearest Neighbour
LDA	Latent dirichlet allocation
LIWC	Linguistic inquiry and word count
LT	Linear threshold
MAP	Mean average precision
MI	Mutual information
NDCG	Normalized discounted cumulative gain
NER	Named entity recognition
NLP	Natural language processing
NYT	The New York Times
OSGs	Online social groups
OSNs	Online social networks
PDLDA	Phrase-discovering Latent dirichlet allocation
PLSA	Probabilistic latent semantic analysis

POS	Part-of-speech
PTM	Probabilistic topic modeling
QA	Question-answer
RG	Reaction gain
RPC	Rate of perplexity change
SCC	Strongly connected component
Sci&Tech	Science and Technology
TF	Term frequency
TNG	Topical n-gram
VADER	Valence Aware Dictionary and sEntiment Reasoner
vs.	versus
WCFP	Weighted categorized frequent posting schedule
WCFR	Weighted categorized frequent reaction schedule
WoM	Word-of-mouth
w.r.t.	With respect to



# Chapter 1

## Introduction

Social networks are currently one of the most popular means of communication and information sharing. Social networking platforms have millions of users, since the introduction of the first social networking platform *Six Degrees* in 1997. Due to proliferation smartphones in recent years, most of the internet users (71%) including celebrities, politicians, commercial organizations have their presence in social networks, and they extensively use social networks for personal and commercial purposes such as marketing, communication, entertainment, to name a few.

Social networks allow users to generate and consume a large amount of information, which play an important role in various tasks such as viral marketing, political campaigns, and job search. As of December 2018, the largest social network Facebook has 2.3 billion monthly active users<sup>1</sup>. Due to huge size of social networks and plethora of generated information, it is difficult to effectively diffuse and summarize the information. We therefore address the problems of information diffusion and information summarization in social networks.

---

<sup>1</sup><https://newsroom.fb.com/company-info/>

## 1.1 Information Diffusion

Information diffusion is a process of disseminating information from an individual or community to another in a social network [1]. Understanding information diffusion in social networks can help to enhance business performance, increase audience engagement, improve personalized recommendation system, and develop a better opinion mining system [2]. Diffusion is usually successful when information reaches to a large number of users in the network. Social networking services allow users to diffuse the information through various reactions such as like, comment, share, and retweet. Business, organizations, and individuals in social networks yearn to increase information diffusion by increasing the number of reactions. A post content with a large number of reactions can increase the visibility of the content, build the reputation of the content creator, and attract other users to give their reactions. Existing studies on information diffusion mainly focus on addressing two questions: (a). how a piece of information spread in social networks [3, 4], (b). how to enhance the information diffusion [5, 6]. The main focus of this thesis is to enhance information diffusion by optimizing the factors that affect information diffusion.

### 1.1.1 Determinants of Information Diffusion

An ample amount of content is generated every day in social media. One of the main goals of content creators is to disseminate their information to a large audience. Many factors affect the information diffusion which includes network connectivity, posting time, post content, location, sentiment, etc. In this thesis, we study three of the most important factors namely, network connectivity, posting time, post content that can highly affect the information diffusion and develop new methods to increase the information diffusion.

## Network Connectivity

A highly connected authoritative user of a social network can diffuse the information widely compared to an ordinary user of the network. To find authoritative users, we can model a social network as a graph, where nodes represent users and edges represent a relationship between users. For example, it can be seen in Figure 1.1, if a few authoritative users (i.e., green nodes) pass the information to other connected users, it can widely spread across the network. Authoritative users can be used to advertise a given topic through word-of-mouth (WoM) marketing as these users get the most attention. WoM marketing is one of the trusted and cost-effective forms of marketing where products are advertised through friends, family, or known authorities. In this thesis, we present a novel algorithm to find authoritative users in online social groups (OSGs) such as Facebook Groups.

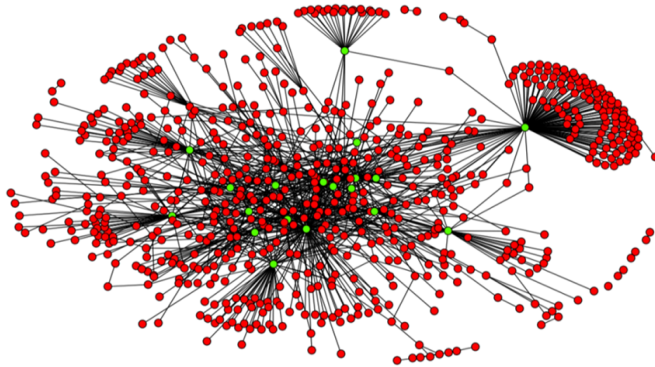


Figure 1.1: Network structure of a online social group

As authoritative position of users varies across the topics [3], it is more useful to find topic-sensitive authoritative users. To this end, we propose a topic-sensitive social interaction graph where edge weights are dynamically computed based on users' interactions and similarity of a post to advertising topic [7]. The interactions could be in the form of likes, likes-on-comment, shares, and comments. To find prominent topic-sensitive authoritative users, we employ link analysis from social network analysis on the topic-sensitive social interaction graph. We also present the concept

of reinforced WoM marketing, where multiple topic-sensitive influential users can together promote a topic or a product to increase the effectiveness of marketing. Finally, to make the marketing most effective, we find the best time to start a campaign in OSGs when a topic can achieve higher user engagement.

## Posting Time

Lifetime of social media content is very short, which is typically a few hours [4,8]. As it can also be seen in Figure 1.2, 50% of reactions are received within four hours of posting on Facebook pages. The main obstacle in getting high information diffusion is that a post has to compete with many other posts within its very short lifetime. If a post content is posted at the time when audience are not online or not interested in interacting with the content, the content will not receive a large number of reactions. Thus a post with less number of reactions will not diffuse to a large audience. On the other hand, a post created at right time can lead to a higher number of reactions and thereby increase the information diffusion. The newsfeed ranking algorithms of social networks also use social interaction counts to determine the rank of the post in the audience feed. In this thesis, we look at the problem of finding the best posting time(s) for a given type of content for it to get a high audience reaction.

For our analysis, we use Facebook pages from five domains, namely e-commerce, traffic, telecommunication, hospital, and politics. To find the best time to post, we derive two classes of schedules: posting-based schedules and reaction-based schedules. The posting-based schedules are computed based on the post creation time. Many admins of Facebook pages may not be aware of when they should post to get maximum audience reactions. However, a few admins with knowledge of Facebook News Feed ranking might have an intuition of when they should post to get maximum audience reactions. We therefore propose three posting based schedules that are based on frequent posting timings [9]. Since our goal is to maximize the number of audience

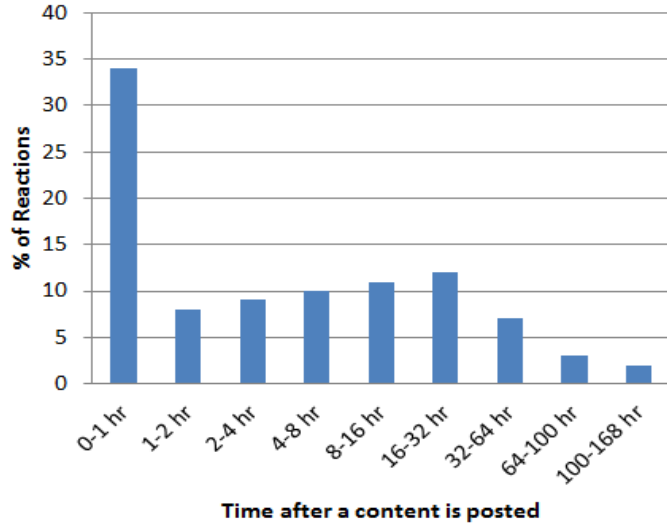


Figure 1.2: Average reactions per post in different time intervals

reactions, the reaction-based schedules are computed based on reaction timings of audience on posts. We analyze the audience reaction timings on created posts and recommend the best time to create a new post for getting a high audience reaction. We propose three reaction-based schedules. The effectiveness of posting and reaction schedules are computed using Reaction Gain. We observe from our experiments in Chapter 4 that Reaction Gain of reaction-based schedules are higher than posting-based schedules. The best reaction-based schedule can give seven times more number of audience reactions compared to the average number of audience reactions that one would get without following any optimized posting schedule. We also determine the type of contents that can increase audience engagement such as videos, photos, etc.

## Post Content

The content of a post plays a crucial role to determine its success or popularity in social networks. Naveed et al. [10] showed that content of a post is one of the most important factors in its popularity. Popularity is measured in terms of audience reactions such as likes, comments, shares, retweets, etc. Among the three popular types of audience reactions in social media, namely likes, comments and shares, interaction in

the form of comments is the most informative. Through comments, users can express their opinions. We say that a post has “high arousal” content if it can attract a huge number of comments from users. In other words, a post has high arousal content if it is on some debatable topic. If a post is getting a huge number of comments, then in social media it will get higher rank because many people would be reading the post and also the corresponding comments. A high arousal content can be useful in various applications such as analyze how a user perceives the news, enhance existing post recommendation systems, increase information diffusion, and understand collective behavior of users’ opinions. In this thesis, we predict high arousal news posts published in social media news pages. We propose an unsupervised approach to label the post of high arousal and low arousal using social interactions [11]. We build two classes of high and low arousal posts. From these posts, we extract multiple features to train an ensemble based voting classifier. Given a new post, we predict whether the post would generate high arousal or not. We also determine the topics of high arousal that can lead to high arousal for a post.

Similarly, a post content with the right sentiment can also increase information diffusion [12]. Sentiment is an important factor of post popularity and it is used in social media to catch the attention of a huge number of users. Due to the lack of inherent regulation, social media is being exploited to spread more aggressive and negative news. In this thesis, we analyze how different types of channels such as TV, radio and print media-based channels use sentiment to garner users’ attention in social media. Same news is presented using different sentiment by different types of channels. Using reaction information on news content, we understand to what extent the sentiment policy employed by social media news channels have been successful in catching users’ attention. Further, comments allow users to express their opinions on a news post. We analyze how users’ opinions depend on the sentiment of a news post and the type of news channel. Existing news recommendation systems can be

enhanced using the sentiment dynamics in social media news channels. Users have preferences of what type of news they like to read. It could be based on the news topic and also the sentiment with which the news is communicated. For example, if someone prefers to read more positive and inspiring news, it would be better to recommend to such users news channels and news topics that mostly have positive sentiment.

## 1.2 Information Summarization

Social media sites generate a huge volume of data every day. Over 90% of web-data is generated in the last two years [13]. The largest social network, Facebook alone generates 4 petabytes data per day<sup>2</sup> and microblogging platform Twitter generates 500 million tweets per day<sup>3</sup>. Due to a massive amount of data generated every day, it is difficult to search, comprehend or categorize the data. Information summarization is a process of creating a concise readable summary from this huge volume of unstructured data. It can be used to generate a summary of user profile, generate a summary of real-world events, improve search and categorization, and create tags for content, individual and community, etc. One of the important approaches of summarization is to find out key topics of interest from a massive amount of data. Topics generated by existing topic modeling methods are good for text categorization but are not ideal for displaying to users because they generate topics that are not similar to manually created topics and are often redundant. In this thesis, we present a novel method to summarize unstructured texts, which uses association mining and natural language processing to generate topics similar to manually created topics [14]. We also show a comprehensive topical summary by grouping semantically related topics using word-embeddings. These summaries can be useful to create gist of real-world

---

<sup>2</sup><https://research.fb.com/facebook-s-top-open-data-problems/>

<sup>3</sup><http://www.internetlivestats.com/twitter-statistics/>

events, generate tags for content, individual and community.

The key contributions of the thesis are as follows:

1. Given a topic, we find topic-sensitive WoM marketers of a social network who can maximize the spread of information in the network. We propose a novel method to create topic-sensitive social interaction graph, which can be exploited to find topic-sensitive WoM marketers.
2. We determine the best time to post that can lead to increase in user reactions on a content. Our best posting schedule can lead to seven times more number of audience reactions compared to the average number of audience reactions that one would get without following any optimized posting schedule.
3. We predict post contents that have a higher potential to generate high arousal, i.e., a large number of reactions, especially comments. We use ensemble-based classifier for arousal prediction and also determine the topics of high arousal that can lead to high arousal for a post.
4. We analyze how TV, radio and print-based news channels use sentiment to get users' attention in their social media pages. We explicate users' reactions to news posts of varying sentiment from different types of news channels. We investigate how news with different sentiments are being perceived by users.
5. We propose a novel topic generation algorithm based on association mining and NLP. As compared to existing probabilistic topic modeling algorithms, our proposed algorithm generates topics that are almost twice more similar to manually created topics of interest and with 13.9% higher precision.

The rest of the thesis is organized as follows. In Chapter 2, we survey the related work. In Chapter 3, we enhance information diffusion through word-of-mouth marketing using network connectivity. Chapter 4 enhances information diffusion by



determining the best time to post to maximize the visibility of content. Chapter 5 presents a novel method to increase information diffusion by posting high arousal content. Chapter 6 enhances information diffusion in social media news channels using sentiment dynamics. In Chapter 7, we present a novel method for information summarization by generating topics of interests. We proceed by concluding this thesis and giving future directions in Chapter 8.

# Chapter 2

## Literature Review

In this chapter, we first present the related work on information diffusion and then review literature on information summarization in social networks.

### 2.1 Information Diffusion

Information diffusion is one of the popular research topics in social network analysis [15–18]. Most of these works address the following two questions: (a) Analyze how information diffuses in OSNs [3, 15, 19], and (b) How to increase information diffusion [5, 16, 20].

To address the above two questions, information diffusion has been modeled using two popular models, namely IC model [21–23] and LT model [24–26]. IC model is an information push/ sender-centric model, where each active user has an opportunity to activate his inactive neighbors with a given probability. LT model is an information pull/ receiver-centric model, where a node is activated by his neighbors if their aggregated weight surpasses his own influence limit. For example, spread of disease in a network can be modeled using IC model whereas spread of opinion in a network can be modeled using LT model. Our work is complementary to existing IC and LT models (refer to Section 2.1.1), where our goal is to address the second question of

how to increase information diffusion in online social networks.

To spread information in a social network, we need to understand the flow and diffusion of information in the social network [27–32]. It requires an understanding of the topological structure and temporal characteristics of the social network [33–37]. In this thesis, we study three important factors of information diffusion, namely network connectivity, posting time and post content. Using network connectivity, we find the influence of users of a social network and use influential users to do WoM marketing. We find the best posting time(s) to enhance information diffusion. We also show how arousing contents and contents with right sentiment can be used to increase information diffusion.

### **2.1.1 Influence of Users in OSNs**

In OSNs, the influence of all users are not the same; it varies based on their interactions and social position. Wu et al. [4] showed that less than 1% of the social network users produce 50% percent of content, while remaining users produce very limited content and have little influence. Vogiatzis et al. [38] stated that news spread by influential users about a product or service may reach up to the maximum possible level. Therefore, it is essential as well as challenging to find a few active authoritative users to widely spread the information within the network.

To discover top influential users or authorities in a social network, we require to understand the topological structure [28,30,32]. Many researchers in this domain have studied the structure of social networks to solve the problem of influence maximization [22,39]. The goal of influence maximization is to maximize product penetration while minimizing the promotion cost by selecting the subset of users that are also called influential users. Further, several studies [40–42] tried to solve the problem of finding influence of a user by using prestige measures (i.e., PageRank [43], HITS [44], Z-score [42], Eigen vector [45]) and centrality measures (i.e., Degree [40], Between-

ness [46], Closeness [47]) from link analysis algorithms. These existing work create a static graph based on user interactions to find the top influential users (or nodes) of the graph. However, they do not focus on finding topic-sensitive influential users. As authoritative position of users vary across the topics [3], we focus on finding topical influential users by creating dynamic graph based on user topical social interactions.

In this thesis, we determine topic-sensitive influential users who can be used to advertise a topic through WoM marketing (refer to Chapter 3). According to Ogilvy Cannes<sup>1</sup>, 74% of consumers identify WoM marketer as a key influencer in their purchasing decisions. According to MarketShare [48], WoM has been shown to improve marketing effectiveness up to 54%. In this thesis, we find top authorities or WoM marketers based on network connectivity and topic-sensitivity to do widespread WoM marketing in a social network. We also propose a concept of reinforced marketing. Marketing would be more effective in giving trust to users about the product if multiple WoM marketers jointly promote the product in a reinforced manner within a network (or group). Similar to LT model, marketing through multiple WoM marketers in the same group help to surpass users' influence limit.

### **2.1.2 Right Time to Post**

A post created at the right time can increase the number of received reactions and thereby increase the diffusion. Users' reaction behavior changes across different social networks [49]. For example, in Twitter, the lifetime of content is quite short compared to other social networks. Some of the topics end in just 20-40 minutes [8]. Wu et al. [4] showed that regardless of the type of content, all contents have a very short life span, which usually drops exponentially after a day. In this thesis, we study the users' reaction behavior in one of the largest social network, Facebook. We reveal that the lifetime of a post content originated in Facebook is also short, and the post receives

---

<sup>1</sup><http://www.adweek.com/prnewser/ogilvy-cannes-study-behold-the-power-of-word-of-mouth/95190?red=pr>

the majority of reactions within a few hours of posting. If a post is not posted at the right time, it may not receive higher user reactions.

To study users' reaction behavior, we need to understand the user dynamics in a social network such as users' connections, users' daily and weekly reaction patterns [41, 50–53]. There have been a few studies on finding the right posting schedule for social network users, which stated that posting time also depends on the user dynamics [54–56]. However, these works are mainly focused on finding the right posting schedule for individual users in a social network. Their posting schedules were derived based on users' social connections and locations. They did not look at many other features that can affect audience reactions, such as features about the content [57–59] or features about the content creator [60].

In this thesis, we attempt to find the right time to spread the information of social media brand pages towards a large audience (refer to Chapter 4). We look at Facebook pages, which has the follower-following type of relationship. A page can have a very large number of followers (e.g., 28 million followers of Amazon page), whereas a user can have at most few thousand friends [61]. We look at a large number of features to find the best posting schedule. In addition to computing schedule for individual pages, we also look at the problem of finding schedule for a group of pages with similar audience reaction.

### **2.1.3 Popularity Prediction of Posts**

Several related works have been carried out to forecast the popularity of social media posts [62–64]. All the existing works in popularity prediction of news posts can be broadly classified into two types: social connection-based methods and content-based methods. Content-based methods can be further divided into two, namely post content-based methods and post sentiment-based methods.

## **Post Popularity using Social Connections**

Social connection-based methods [65–68] use social features such as number of friends, followers, etc., to predict content popularity. Zaman et al. [65] studied reaction behavior of retweetability among users. They used author information such as number of followers, identity of the source, etc. Suh et al. [66] analyzed the factors that impact retweeting and showed that the number of followers and friends have a lot of impact, while factors such as number of statuses and favorites do not. Petrovic et al. [67] used passive-aggressive algorithm to predict if a tweet will be retweeted, which would lead to high information spread through a large number of users. Weng et al. [68] predicted the future popularity of an article using its early spreading patterns. They concluded that features based on community structure are the most powerful popularity predictors.

## **Popularity using Post Content**

There are several content-based methods to predict the popularity of news posts [10, 64, 69, 70]. Bandari et al. [71] predicted the popularity of posts using post contents. They considered four features, namely category of the news post, subjectivity of the news post, named entities present in the news post, and source of the news post. They used both classification and regression to predict popularity. Lee et al. [69] proposed a framework that can predict the number of comments by observing an article for 2-3 days. Tatar et al. [64] proposed a method to rank the news posts by predicting users comments.

In contrast with existing works, we perform our analysis on social media news channels where the goal is to predict the arousal of news posts before it is published in social media (refer to Chapter 5). Arousal is similar to popularity with the difference being that, arousal ensures lots of user feedback or comments on a post. We derive the features from news articles that are related to news coverage and popularity and

select the features related to arousal prediction using word-embeddings.

## **Sentiment Dynamics of News Posts**

Sentiment of a post is one of the important factors of high information diffusion and popularity. Naveed et al. [10] showed that negative news is more attractive to users and easily catch their attention. They used 15 different set of content-based features and predicted the likelihood of a tweet being retweeted using logistic regression. Wu et al. [70] showed that the lifetime of negative news is very short but positive news stay for a longer time. They predicted the decay of social media content using classification techniques. However, in this thesis, we show that it is not only the negative news that catch user attention but also positive news can garner a lot of user attention (refer to Chapter 6). Popularity gained by both positive and negative news is usually higher than neutral news.

Researchers [72] at Universidade Federal de Minas Gerais (UFMG) developed a tool that present news to users based on their interest or polarity. They ranked news articles based on their popularity and sentiment score. Reis et al. [73] analyzed the news headlines from a popular global news channel. They showed that sentiment of headline correlates with the popularity of news and negative comments are posted independently of the sentiment score of the headlines. On the other hand, our analysis shows that although the sentiment polarity of a news post correlates with the popularity of posts, the polarity of a comment is not completely independent of the polarity of the actual news post; it is a function of the polarity of the news post.

Zubiaga [74] explored the problem of finding the topics of interests of users in social media and a news channel, namely The New York Times. He showed that the top topics created by the newspaper are mainly related to hard/big news such as Politics, Money, World, etc., whereas users are more interested in posting niche news in social media. However, in this thesis, news posts are created by social media

channels and not by news readers. Common people would be more interested to post niche news in social media, compared to big news/headlines, as niche news draws more attention. In contrast, for news channels both the types of news are important, and we found their social media pages to cover a wide range of news.

## 2.2 Information Summarization

With the ever-growing volume of user-generated content, it is becoming difficult to summarize these contents. Multiple methods are proposed to summarize a large corpus of these contents using snippets [75,76], hashtags [77,78], word-cloud [79], and key-topics of interest [80,81]. A document corpus can be summarized either using top representative sentences or finding the best concepts (or key-topics) from the corpus. A few sentences may not provide different hidden or latent topics of the corpus. So, it is useful to summarize a document using key-topics to get diverse themes or topics, which could help in better understanding and categorization of the corpus.

### 2.2.1 Key-topics of Interests

Many statistical methods [80,82–85] have been proposed to determine topics for text documents. One such topic model, Latent Dirichlet Allocation (LDA) [80] that relies on bag-of-words assumption has a great impact in the fields of machine learning and text mining. LDA considers the document as a mixture of topics, and the topics are multinomial distribution of words present in the document. Unlike PLSA [86], the LDA model is a well-defined generative model and can create topics without overfitting. However, LDA does not generate meaningful phrases. It generates many generic words that do not convey complete information. For example, LDA generates many trivial words such as ‘model’, ‘data’, ‘algorithm’, ‘approach’, and ‘control’ with high term-topic probability from the Machine Learning publications.



To identify phrases from text documents, several methods have been proposed [81, 87–89]. Two such methods are Topical N-Gram (TNG) and Phrase-discovering Latent Dirichlet Allocation (PD-LDA) that generate phrases. These methods generate phrases but often suffer from high complexity and poor scalability. Moreover, topics generated by existing topic modeling methods are good for text categorization but are not ideal for displaying to users because they generate topics that are not so readable and are often redundant. In our experiments, these methods generate many less interpretable phrases such as ‘empirical study’, ‘based malware’, ‘case study’ which cannot be used to label the topics of interest of research communities such as research conferences and research areas.

In this thesis, we generate a concise summary from a large amount of social network research publications using key-topics of interests (refer to Chapter 7). We propose a method based on association mining and NLP that generates more meaningful topics. We present several NLP based refining rules to get well-formed topics that are similar to manually created topics. Results of our evaluations show that the proposed method generates topics that are more interpretable and meaningful than those generated by existing methods.

# Chapter 3

## Information Diffusion through Topic-sensitive WoM Marketing

### 3.1 Introduction

Network connectivity is one of the most important determinants of information diffusion. A highly connected authoritative user of a social network can diffuse information widely compared to an ordinary user of the network. As social networks are one of the biggest sources of information sharing and communication, network connectivity can be used to enhance business performance through social marketing. In this chapter, we use network connectivity of a social network to enhance information diffusion through topic-sensitive word-of-mouth (WoM) Marketing.

Marketing is a process by which products and services are introduced that have utility for customers, sellers, and society. Some of the major goals of the marketing are to increase the revenue, build the reputation of a company, and maintain healthy competition. In order to implement effective marketing, one has to identify the target customers, understand their needs and execute the most effective marketing method. Many marketing methods exist such as field marketing, B2P marketing, direct mar-

keting, online marketing, WoM marketing, etc. With the emergence of the Internet, online marketing has become one of the biggest sources of marketing. In online marketing, advertisements provide a range from basic text descriptions with links to rich graphics with slideshows. However, a major problem with the most of online advertisements is that people have a lack of trust on these information sources. Also, people receive a large number of online advertisements daily, which have made them immune to these advertisements. These problems can be diminished if a product is advertised through WoM marketing. In WoM marketing, information is passed by a set of trusted or known people. According to Whitley<sup>1</sup>, 64% of marketing executives report that they believe WoM marketing is the most effective form of marketing. The reason is that people believe on the words of the known people such as friends, family and closely known authorities.

WoM marketing through only friends or family will be in a limited domain and quite restricted. Since people use social media regularly to access different types of information, in this chapter we propose the use of social media to do widespread WoM marketing. Different types of social network models exist, such as friend-to-friend, question-answer, and follower-following social networks. We exploit question-answer (QA) type of network to do WoM marketing. There are many QA type of networks, such as Stack Overflow, Quora, Reddit, online social groups, etc. In this chapter, we utilize online social groups (OSGs) such as Facebook groups to show how one can use these groups to perform widespread WoM marketing. The members of a Facebook group have more focused interest compared to a generic friend or follower-following networks. For example, Data Science, C/C++ programming, Java for developers, etc., are some very popular and focused public groups.

Facebook groups are concentrated on a specific topic and many of them have a large number of members. Businesses can use prominent and reliable members of

---

<sup>1</sup><http://www.forbes.com/sites/kimberlywhitley/2014/07/17/why-word-of-mouth-marketing-is-the-most-important-social-media/#7762b8f07a77>

such groups to do marketing. Active members of groups often help other members by posting useful pieces of information in the form of posts and comments, and in return they get publicity in the form of user reactions, such as likes, comments, shares, from other members. In due course of time, these active members gain the trust of other members of groups and become influential users. Businesses can use these trusted influential users of groups to market their products by giving them incentives. Since a topic is advertised by one of their trusted peers, with influential position, members of the group pay more attention to such recommendations. To make the problem more tangible, consider the following example:

**Example:** Consider a book publisher having a limited advertisement budget wants to advertise a Java book in a Facebook group. The publisher would like to identify a few users who can promote the book to a large number of users. The publisher can target influential users by giving some free sample copy or discount. To this end, the publisher would be interested in knowing the answers to the following questions:

1. Who are the top- $k$  influential users in a Facebook group?
2. Given an advertising topic, what fraction of the group would be influenced by a selected set of influential users?
3. How to implement reinforced marketing so that each topic is marketed collectively by at least  $k$  influential users?
4. What are the best periods of a year when marketing would be the most effective?

In this chapter, we answer all the above questions. Our key contributions are as follows:

- We present a graph-based method to find topic-sensitive influential users in OSGs for WoM marketing.

- We propose a novel method to create a dynamic topic-sensitive social interaction graph from users' interactions in a group.
- We exploit dynamic graph to find influential users using link analysis from social network analysis.
- We finally analyze the effectiveness of marketing in OSGs across different topics and time periods.

The rest of the chapter is organized as follows. In Section 3.2, we formally define the problem of information diffusion through topic-sensitive WoM marketing. In Section 3.3, we study the network structure of OSGs. In Section 3.4, we present the novel method to create topic-sensitive social interaction graph. In Section 3.5, we describe how to find top influential users in OSGs. In Section 3.6, we propose the concept of reinforced marketing. We proceed by describing the experimental evaluation and the results in Section 3.7. Finally, we conclude our work in Section 3.8.

## 3.2 Problem Definition

We define the problem of information diffusion through topic-sensitive WoM marketing in terms of the following sequence of sub-problems:

**Problem 1 (Create Social Interaction Graph):** *Given a topic  $T$ , a Facebook group  $F$  and the interactions  $I$  in the group  $F$ , create a social interaction graph  $G(V, E)$ , where the vertices  $V$  represent the members of the group  $F$  and edges  $E$  represent the interaction between group members.*

Problem 1 is to create a social interaction graph for a given Facebook group. The interactions in a group include creation of posts and reactions to posts, such as likes, comments, likes-on-comment, and shares. The members of the group represent vertices and interaction among members (users) represent edges of the graph. It is

a topic-sensitive graph, where edge weights are dynamically computed based on the given topic  $T$ . We assign a weight to the edge based on the given topic  $T$  and reactions that a user had done to the post or comment created by another user.

**Problem 2 (Finding Influential Users):** *Given a social interaction graph  $G(V, E)$  and a topic  $T$ , find the top- $K$  influential users  $U$  from the graph  $G(V, E)$ , who can give maximum visibility to the topic  $T$  in the corresponding Facebook group of the given interaction graph  $G(V, E)$ .*

People form Facebook groups to explore about certain topic. Naturally, in such groups some members with more knowledge become authorities, whose words have great influence on the other group members. In this problem, our goal is to find influential users for a given marketing topic.

**Problem 3 (Reinforced Marketing):** *Given a social interaction graph  $G(V, E)$ , a topic  $T$  and reinforcement parameter  $r$ , find the set of influential users  $U_R$  such that the marketing of each influential user from  $U_R$  can be reinforced by at least  $(r - 1)$  other influential users.*

The social position of a user has an important effect on marketing. If someone who is not an authority markets a product, the marketing will hardly have any impact. If one authority markets the product, the marketing will be more effective. The marketing will be even more effective if multiple authorities can collectively market the product. When people hear the same message reinforced by multiple authorities that they trust, it is more likely they will consider buying the product. Thus, we need to find authorities in such a way that if one authority markets the product, there are at least  $(r - 1)$  other authorities in the set  $U_R$ , who can support the marketing. These  $(r - 1)$  other authorities should be closely related with other members whom the above mentioned one authority will market.

### 3.3 Analysis of Online Social Groups

Only a few members of a group usually have a significant influence on the group [40]. These members of the group post most of the information and rest of the members consume the information in the form of views, clicks, likes, comments, shares, etc. To get a better insight into users' social interactions in OSGs, we do two types of analysis: structure analysis using bow-tie structure and connectivity analysis using degree.

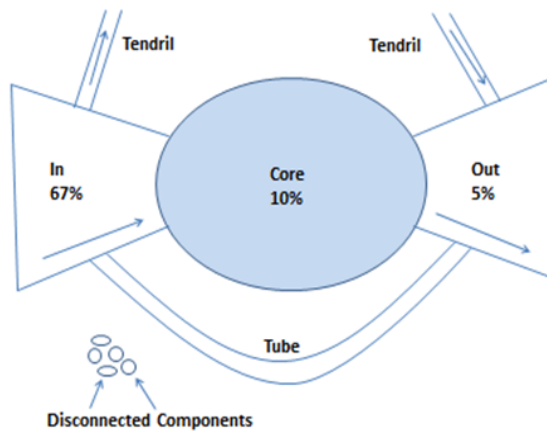


Figure 3.1: Bow-tie Structure of OSGs

We use bow-tie structure [90] to examine the general structure of OSGs. It has five distinct components namely *core*, *in*, *out*, *tendrils* and *tube*. *core* is a strongly connected component (SCC) and contains users who often help or interact with each other. The *in* component includes users who only react to contents. The *out* contains users who only post contents. *Tendrils* and *tubes* consist of users who connect to either *in* or *out* or both but not to the *core*. *Tendrils* users only react to contents created by *out* users or whose contents are only reacted by *in* users. *Tubes* users connect to both *in* and *out*.

We perform structural analysis on 100 Facebook Groups (or OSGs) having more than 20,000 members. We show a bow-tie structure of groups in Figure 3.1. We observe that OSGs have much bigger *in* component compared to *out* and *core*. This

indicates that in OSGs, most of the users (67%) only react to contents, and 5% of users only post contents. There are only 10% core users who post contents as well as react to the contents of each other. These results show that OSGs are information seeking communities where most of the users consume information, and very few users generate information. Most of the users join a group to keep themselves updated by getting information related to topics of shared group interest.

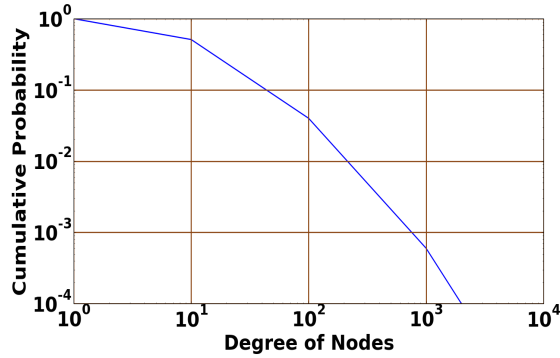


Figure 3.2: Degree distribution in groups

We next do degree analysis to get insight into the connectivity of users' in OSGs. Degree is a general way to reveal users' relative connectedness in a large complex network. The degree distribution shows the number of users (cumulative probability of users) in a network with a given degree. As we can see in Figure 3.2, the degree distribution appears to follow the power law. Most of the users have very less degree, which indicates that these users are connected to just a few other users. However, very few users are connected to a large number of users. As these few users have a large number of connections, they can effectively spread the information to a large number of users.

### 3.4 Social Interaction Graph

In this section, we give a solution to Problem 1. We create a topic-sensitive social interaction graph based on social interactions in a group. To create the topic-sensitive



social interaction graph, we first compute the topical relevance of group users and then construct a graph based on their topical interactions.

### 3.4.1 Measuring Topical Relevance

To measure the topical relevance of users, we analyze the content of their posts. We first find the users who are interested in a given topic  $T$ . Based on the fact that users who have posted the content related to topic  $T$  would be interested in topic  $T$ , we select influential users from these interested users to market topic  $T$ . For example, to market a *Database* book, we need to find the users who have posted contents related to *Database*. One simple approach is to look into all the posts, which contain the word *database* in them. This approach, however, fails to give good results as there might be some posts, which are actually relevant to the *Databases* but do not contain the word *database*.

In order to effectively identify relevant posts, we generate a list of words which are semantically related to the given seed word or topic  $T$ . For example, some of the important words related to *database* are *sql*, *query* and *schema*. The words related to a topic may have a different degree of affinity or relatedness. For example, word *sql* is more closely related to *database* as compared to word *query*. For this task we need a system which, given a word, gives a list of relevant words along with its relevance score.

In this chapter, we use *Semantic Link*<sup>2</sup> system that gives a list of semantically related words for a given seed word. It uses the fact that a word can be recognized by the association that it keeps [91]. It indicates that similar words often occur together. For example, words *database* and *sql* often occur together. These are semantically related words, meaning that their co-occurrence is not by chance but rather due to some non-trivial relationship. Such relationships include similar syntactic rules or

---

<sup>2</sup><http://semantic-link.com/>

similar meanings. *Semantic Link* attempts to capture such relationships between words and uses these relationships to find related words.

*Semantic Link* analyzes Wikipedia and finds all the words or topics, which are semantically related to a given seed word. It uses Mutual Information (MI) [92], which is a measure of the mutual dependence between two topics. Higher the MI score for a given pair of topics, higher the chance that they are related. MI score is defined as follows:

$$MI(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (3.1)$$

where  $MI \in [0, 1]$ ,  $X$  and  $Y$  are two random set of topics.  $p(x, y)$  is the joint probability distribution function of  $X$  and  $Y$ .  $p(x)$  and  $p(y)$  are the marginal probability density functions of  $X$  and  $Y$  respectively.

After getting a list of related words, we find posts relevant to these words. One approach is to filter out the posts that do not contain any related words. However, this approach has a limitation that users who do not have relevant posts (related words in their posts) will have no connection with users who have relevant posts. Such users thus will not contribute in determining the rank (or influential position) of relevant users. In such cases, only the popularity of relevant users will matter while determining ranks of users. As this approach ignores the relationship of relevant users with non-relevant users, a better approach is to give higher weight to the relevant posts and their interactions. The weight is computed based on the presence of related words in posts. To this end, we first compute *relevance* score for every post, which is the sum of MI score of all the related words that are present in the post. We next compute boosted relevance (*bRelevance*) based on the *relevance* score. Boosted relevance is the most important determinant of the weight and is computed as follows:

$$bRelevance = 1 + \varphi \times \log(1 + relevance) \quad (3.2)$$

where  $\varphi$  is a topic-sensitive user parameter in our system, which controls topic-sensitivity in the graph. If  $\varphi$  is too high, the graph will be highly topic-sensitive. And if  $\varphi$  is 0, the graph will be a general social interaction graph that does not capture the topic-sensitivity. Similarly, we compute  $bRelevance$  for textual comments based on the relevant words present in it and assign a higher weight to the relevant comments and their interactions.

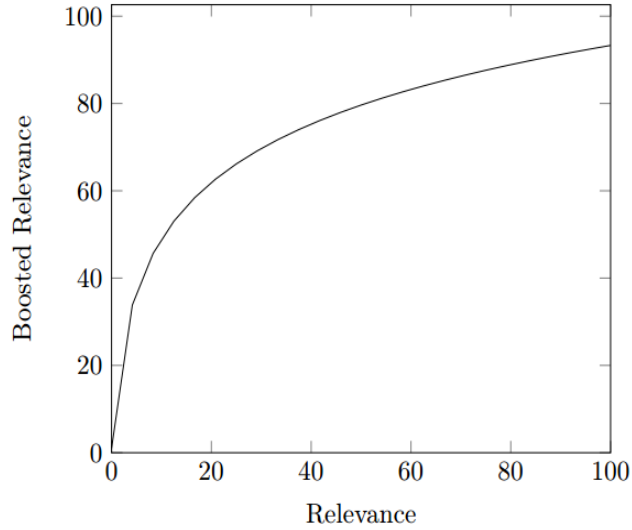


Figure 3.3: Boosted relevance

Figure 3.3 shows the effect of the boosting function. The slope of the graph decreases as relevance increases. It indicates that if a user tries to spam the system with too many words related to advertisement topic, the logarithmic function of boosted relevance is not increased too much. Further, Algorithm 1 presents the method that we use to compute the  $bRelevance$  of different posts. Lines 3 to 7 show how to compute the semantic similarity ( $relevance$ ) between topic word ( $tWord$ ) and post word ( $pWord$ ). Line 8 shows the equation to compute the boosted relevance using relevance score.

---

**Algorithm 1** Algorithm for Computing Boost to Interactions

---

*Input:*  $T$ : set of topic words  
 $P$ : Post  
*Output:*  $bRelevance$ : boost of post  $P$   
*Method:*  
1:  $relevance \leftarrow 0$   
2:  $postWords \leftarrow P.getWords()$   
3: **for all**  $tWord \in T$  **do**  
4:     **for all**  $pWord \in postWords$  **do**  
5:          $relevance+ = Similarity(tWord, pWord)$   
6:     **end for**  
7: **end for**  
8:  $bRelevance = 1 + \varphi * \ln(1 + relevance)$   
9: **return**  $bRelevance$

---

We derive a graph structure from the group by representing each user of the group as a vertex of the graph and each user interaction such as *like-on-comment*, *like*, *comment*, *share* as an edge of the graph. We create an edge from user  $u_i$  to  $u_j$ , if the user  $u_i$  has reacted to any post or comment that is created by the user  $u_j$ . The weights of the edges are dynamically determined by the product of the weight corresponding to the type of interaction with the boosted relevance.

We give different weight to different types of interactions. As suggested by Bucher et al. [93], ‘share’ and ‘comment’ require higher commitment or exertion than ‘like’. Therefore, ‘comment’ and ‘share’ outweigh ‘like’. Further, ‘share’ generates a higher amount of engagement compared to ‘comment’ as a shared post appears on user’s profile page. A shared post is further pushed towards user’s connections as it constitutes a part of user’s self-presentation. This indicates that ‘share’ outweighs ‘comment’. ‘like’ outweigh ‘like on comment’ as like on a post provides higher visibility to the actual post content. In our experiment, we empirically set the value of ‘like on comment’, ‘like’, ‘comment’, ‘share’ to 1, 2, 4, 8 respectively.

## 3.5 Finding Influential Users in OSG

In this section, we describe how to find influential users in OSGs. This is a solution to problem 2.

We use link analysis algorithms from social network analysis to find influential users [94]. The two primary types of link analysis algorithms are prestige measures [42–44] and centrality measures [40, 46, 47]. Centrality measures focus on out-links whereas prestige measures focus on in-links while finding prominent users in a network. We use one of the prestige measures, PageRank [43] to find the top influential users. One of the reasons to use PageRank is that unlike other link analysis algorithms [42, 47], it considers the importance of each user who interacted with a target user to determine the influential position of the target user.

PageRank was originally developed to rank the web pages for search results. Web pages are connected together through hyperlinks. Similarly, we have a topic-sensitive social interaction graph where users are connected through social interactions. We therefore apply PageRank algorithm on the social interaction graph to find influential users. We rank users based on decreasing order of PageRank score and select the top- $k$  users to be the potential WoM marketers. We also compare the effectiveness of PageRank algorithm on social interaction graph with other link analysis algorithms such as Z-score [42], Eigen vector [45], HITS [44], Betweenness [46] and Closeness [47]. We use all these link analysis algorithms as authority measure algorithms to find top authoritative or influential users.

## 3.6 Reinforced Marketing

In this section, we give a solution to problem 3. In OSGs users interact with other users having similar topics of interests. This limited user interaction leads to the formation of sub-groups. Since user interactions may get confined to a few sub-

groups, it is important to find multiple influential users from each sub-group so that they can collectively promote the product, which will be more effective in giving trust to users about the product. We need to do this for all the important sub-groups.

We find sub-groups by determining weakly connected components in the graph. A weakly connected component is a maximal sub-graph of a directed graph such that for every pair of vertices in the sub-graph, there is an undirected path. Each member of a weakly connected component would have reacted on someone’s post in the group or would have received a reaction from someone else in the group. To perform effective marketing, we target a sub-group only if it contains enough users. If users in a sub-group are less than threshold  $th$ , we do not select that sub-group for marketing.

We apply the best authority measure algorithm (described in Section 3.7) in a topic-sensitive social interaction graph to find the top- $k$  topic-sensitive influential users of the group. For each of the sub-group, we select top- $r$  ( $r < k$ ) users from the set of  $k$  users such that these  $r$  users also belong to the same sub-group. These  $r$  influential users can support each other by advertising the same product to their sub-group(s).

## 3.7 Evaluations

In this section, we first give details about our dataset and evaluation metrics. Later, we compare the performance of various authority measure algorithms and show the characteristics of influential users through some anecdotal examples.

### 3.7.1 Experimental Setup

We extract the dataset from Facebook groups for the experiment as these are focused groups with a large number of users. Facebook groups are communities of people

where they share their common interests in the form of posts and comments. Members of groups can react to the posts/comments created by each other. Reactions consist of a textual comment and a unary rating score in the form of likes and shares. We use Facebook Graph API [95] to collect the dataset. The dataset contains 100 of Facebook groups having at least 20,000 members. It includes 0.3 million posts and 10 million reactions that were created in 5 years (from 2011 to 2015). We perform various text pre-processing tasks on textual contents of the dataset such as stop words removal, stemming and lemmatization [96].

### 3.7.2 Evaluation Metrics

We show the effectiveness of algorithms by using three metrics, namely correlation, precision, and influence.

**Correlation:** We use correlation metric to measure the strength of association between two ranks. We use Pearson Correlation [97] to evaluate the effectiveness of authority measure algorithms. We find the correlation between rank assigned to users by authority measure algorithms and the baseline influence metrics (described later in this section).

**Precision and Normalized Discounted Cumulative Gain:** We use these metrics to check the quality of authority measure algorithms by computing the relevancy of top- $k$  influential users generated by these algorithms. Precision is a fraction of retrieved users that are influential. Normalized Discounted Cumulative Gain is computed based on the Discounted Cumulative Gain [98], which includes the position of users in the consideration of their importance.

**Influence:** We use two influence metrics as baselines to evaluate the user’s authority position in OSGs, namely centrality and popularity measures. Degree is a centrality measure that evaluates the user’s connectivity whereas votes and topical votes are

popularity measures that evaluate the user’s prestige. For each user, we compute votes by taking the weighted sum of all the reactions received by the user over all his posts and comments. However, we compute topical votes by taking the weighted sum of reactions over all his posts and comments that contain the advertisement topic itself or the topics semantically related to the advertisement topic.

### 3.7.3 Effectiveness of Algorithms

User behavior dynamics in OSGs is different from other networks. An algorithm that works efficiently for one network may not be appropriate for OSGs. Therefore, it is essential to evaluate the effectiveness of authority measure algorithms in OSGs. To evaluate the effectiveness of authority measure algorithms, we use Pearson Correlation metric. We compute the correlation of top-200 influential users generated by authority measure algorithms with the votes received by these users. For an authority measure algorithm, we compute the aggregated correlation by averaging the correlation across all the groups.

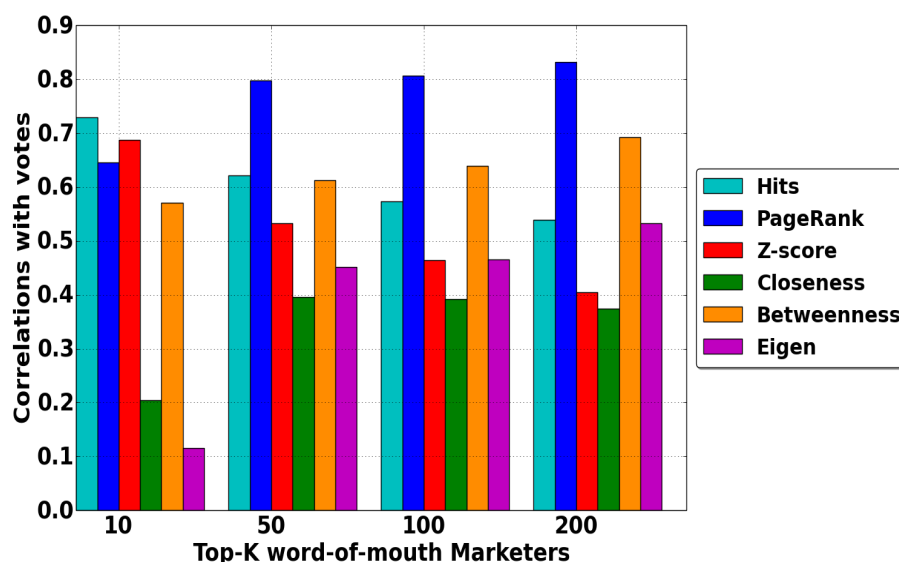


Figure 3.4: Correlation of authority measure algorithms with votes



Figure 3.4 shows the correlation of the top-200 word-of-mouth marketers (or influential users) ranked by the various authority measure algorithms with the votes. As can be seen in the figure, prestige measure HITS and PageRank outperforms the other algorithms. HITS performs better than other algorithms for top-10 users whereas PageRank outperforms for top-50 or more users. One of the reasons is that PageRank is a global measure and it does not trap in local neighborhood. However, HITS usually suffers from topic drift [99]. Further, Betweenness tends to produce slightly better results than most of the other algorithms. One of the reasons is that nodes having high Betweenness are the bridges of two parts of the graph (sub-graph) and have the potential to disconnect the graph if removed. If a user having high Betweenness posts an update, there is a high chance that it will spread rapidly across different sub-graphs.

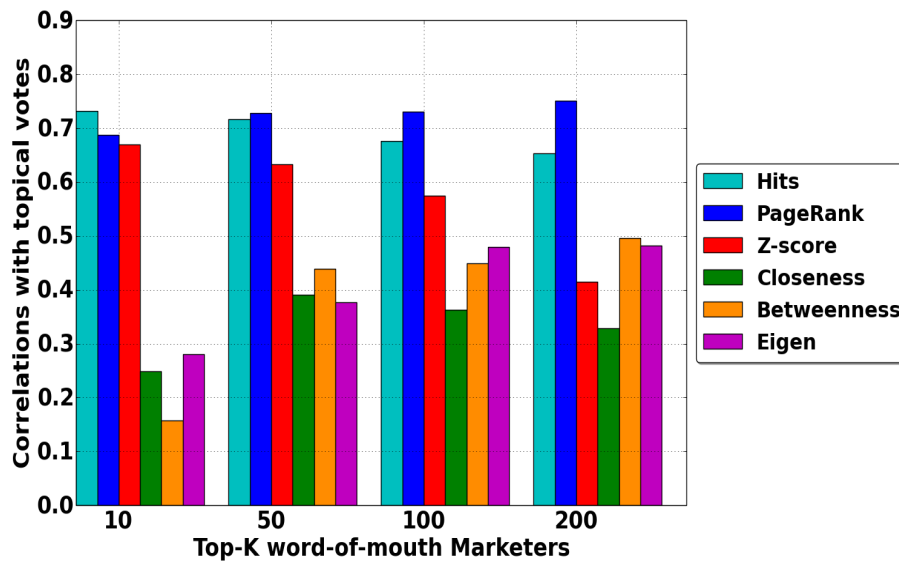


Figure 3.5: Correlation of authority measure algorithms with topical votes

Figure 3.5 shows the correlation of the top-200 influential users ranked by the various authority measure algorithms with the topical votes. HITS performs better for top-10 users whereas PageRank performs better than HITS for top-50 or more users. In conclusion, PageRank can be utilized to find influential users for general

marketing as it shows high correlation with both votes and topical votes.

### 3.7.4 Precision Analysis

We evaluate the correctness of authority measure algorithms by using Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG), which are standard measures to evaluate the effectiveness of web page ranking algorithms.

We consider top-50 influential users of the groups generated by algorithms. We asked five students of our research lab to join these technical groups and manually judge whether a user is influential or not from their viewpoints for a given topic  $T$ . We also asked to rank these users for a given topic. We provided all the posts and reactions of influential users to the students. These students labeled the data independently, and percentage of agreement among the students was 92%. We use this label data as ground truth for finding MAP and NDCG of the algorithms. We compute the overall MAP, NDCG by averaging MAP, NDCG of all the groups respectively.

<b>Authority Measures</b>	<b>MAP</b>	<b>NDCG</b>
PageRank	0.91	0.83
HITS	0.87	0.75
Z-score	0.70	0.65
Eigen	0.72	0.69
Betweenness	0.76	0.70
Closeness	0.73	0.67

Table 3.1: MAP and NDCG of authority measure algorithms

As can be observed in Table 5.1, for a given topic  $T$  PageRank performs better than other authority measure algorithms. PageRank finds topic-sensitive influential users with the highest MAP and NDCG. We therefore use PageRank for our analysis in the rest of the chapter.

### 3.7.5 Marketing Across Topics

In this section, we study the behavior of influential users across different topics and investigate how widely the rank correlation of these users changes by changing advertising topics.

Top influential users (or top users) for all the topics are not different. Top users tend to express their opinions on many popular topics of the group. To inspect the dynamics of top users' across different topics, we compare the relative order of their ranks across topics. We ignore the least popular topics and focus on the set of relatively popular topics. We apply Topical n-gram [81] on the posts to find popular topics of the *Java For Developers*<sup>3</sup> group (Java group). Web, Servlet, and Constructor are some popular topics in the group. We choose these topics to measure the variation in top users ranking across these topics. We use correlation to compare the ranking patterns of top users for pairs of topics.

<b>Topics</b>	<b>Top-20 Users</b>	<b>Top-200 Users</b>
Web vs. Servlet	0.79	0.56
Web vs. Constructor	0.53	0.46
Constructor vs. Servlet	0.49	0.39

Table 3.2: Correlation in top users ranking for popular topics

As can be seen in Table 3.2 that correlation is high for the top-20 users, which indicates that these users post over a wide range of topics. Among topic pairs, {Web, Servlet} shows the highest correlation for the top-20 users. This is because these two topics are closely related to each other. Servlets are used in ‘web programming’ and ‘web designing’ applications. This analysis indicates that top users hold significant influence over a range of topics and can be used to spread information related to popular topics of OSGs.

To get more insight into variation in correlation of top influential users across

---

<sup>3</sup><https://www.facebook.com/groups/java4developers/>

topics, we perform the experiment on wide range of topics in Java group. We select 20 topics from each of popular topics, less popular topics, and unpopular topics. We compute Mutual Information (MI) score for all these topics with respect to group topic (shared group interest or common interest of the group). We derive top-200 topic-sensitive influential users by using these topics and measure the correlation of these users with topical votes.

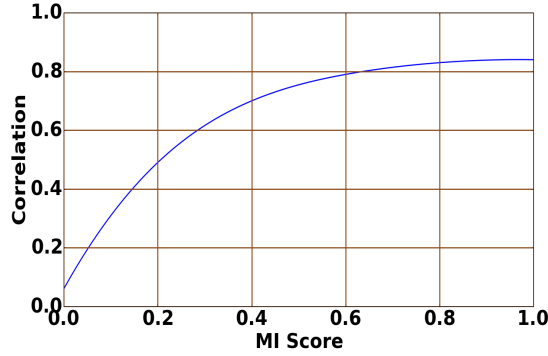


Figure 3.6: Correlation of top users across variety of topics

As it can be observed in Figure 3.6 that correlation decreases as MI score decreases. If a chosen topic has very less similarity with the group topic, then authority measure algorithms show very less correlation. It signifies that the quality of top users also depends on the topic. If an advertising topic is less related to shared group interest then it is less likely to get prominent topical users who can influence the whole group as the quality of these users decreases. Therefore, it is recommended that advertising businesses should select a topic, which is highly related to shared group interest to do effective marketing in OSGs.

We next understand through an example that how the rank of an influential user changes by changing the advertising topic. We perform this experiment in a relatively smaller group. One of the groups that we analyzed is related to cooking called *Whats Cooking*<sup>4</sup>. In this group, the people discuss the recipes of various food items. We

<sup>4</sup><https://www.facebook.com/groups/308194239304304/>

extract the posts and their reactions from the group. We find the top-20 influential persons related to the topic ‘egg’. We look into the recent posts of the most influential user (user A) tagged by our algorithm. We find that her recent posts are actually related to ‘egg’ and are appreciated by several other members of the group. One of her recent posts is as follows:

*Garlic and Sour Cream Scrambled Eggs, Broiled Boudin and Grilled Muenster and Mexican Cheese Toast Squares. Gmorning Ya’ll!!!!*

Next, we find the top-20 influential users for the topic ‘cake’. The most influential user for ‘cake’ is found to be another user (user B). User B got 3<sup>rd</sup> rank in the previous list of influential users related to ‘egg’. After checking the content posted by user B, we find that she has posted quite a lot about cakes recently. One of her posts with high visibility has the following text:

*Ladies and Gentleman I present to you Tia’s 7-UP Pound cake! Think it looks pretty good. Tell yall how it tastes tomorrow!!!!*

There is one point common between these two users that they both are popular users of the group. They post lots of content in the group. These results show that the proposed algorithm works well and gives appropriate WoM marketers for different topics.

### **3.7.6 Empirical Evaluation**

In order to investigate influential users’ characteristics and behavior dynamics, we find the connectivity of influential users and their structural position in OSGs. First, we find indegree connectivity of top influential users in the *Java For Developers* group (Java group) having 35,000 members at the time of the experiment. We observe that average indegree of top-20 users is 1604 whereas average indegree of the whole group

is 8. The reason for this is that authority measure algorithm strongly correlate with the indegree of the top users.

To get more insight into the structural position of influential users in the group, we present the network structure of influential users, which is constructed in a similar way as mentioned in Section 3.4. We take a small instance of Java group with 707 nodes, 1187 edges and present a network structure of the Group in Figure 1.1. The users of the network can be divided into two types: top users and ordinary users. The green color nodes represent the top-20 users, and the red color nodes represent ordinary users of the group. We observe in Figure 1.1 that top users are strongly connected with the large number of members of group. Statistics reveal that average degree of the group is 3.35 whereas average degree of top-20 influential users is 72 in this small instance of Java group. Moreover, average number of reactions received by a user of the group is 5.2 whereas average reactions received by top-20 influential users is 98. These statistics reveal that top influential users are connected to a large number of users of the group and receive a large number of users' reactions.

### 3.7.7 Temporal Dynamics

We analyze posting and reaction behavior of top influential users over a period of time and find the right time to start the promotion in a group to maximize content visibility. Our results are based on five years of temporal data.

In order to examine the influential users' posting behavior, we select the top 1000 influential users based on their ranks in the Java group. We divide top users into three groups based on their ranks such as top 200 users, top 201-500 users, and top 500-1000 users. We aim to analyze the differences in posting behavior of these users. We compute the probability of posting a post for all these three groups in each month of the year. Figure 3.7 shows the time evolution of the posts of influential users (top users).

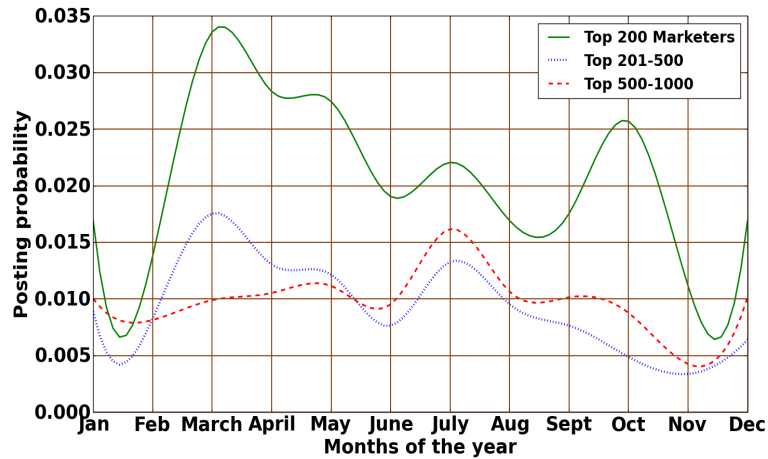


Figure 3.7: Posting behavior of top users

Our findings about the posting behavior of top users reveal two interesting observations. First, top users post significant updates over a period of time. Top 200 users post lots of information compared to top 500 and top 1000 users. Second, lots of posts are posted during the month of March, April, and October. This is perhaps due to various competitive and semester exams in India during these months, which motivates top users to post a lot of information about various topics. So, it is better to choose these periods of a year for marketing.

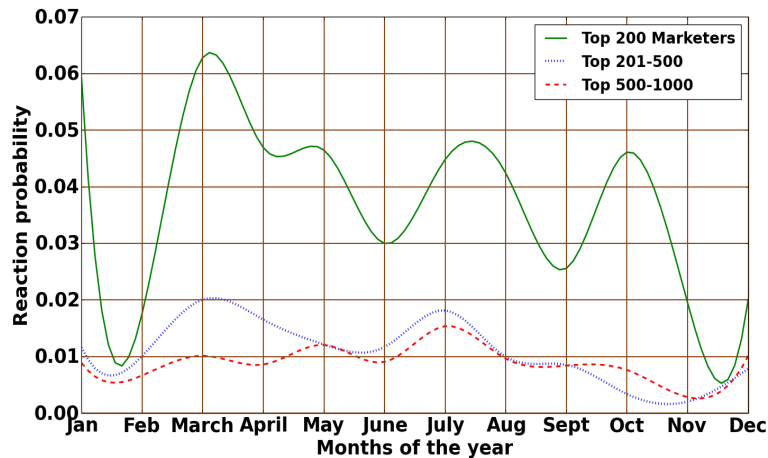


Figure 3.8: Reaction behavior of top users

We also perform a similar experiment on reactions received by top users. Users of a group react to contents created by other users of the group. We do an analysis to

investigate how users of the group react to contents created by an average user and the influential user of the group. As can be seen in Figure 3.8, reaction pattern follows the same trend as posting pattern, i.e., more number of user reactions in the month of March, April, and October. It is due to a large number of posts created by top users during these periods of months and this posting behavior leads to increase the number of user reactions. As lots of users are active during these periods of months, advertising companies can target more number of top users to promote products during these periods.

### 3.8 Conclusion

In the chapter, we proposed a method to enhance information diffusion through WoM marketing. We presented a novel algorithm to create topic-sensitive social interaction graph from users' interactions in the network. We used link analysis algorithms on social interaction graph to find topic-specific influential users. Organizations can promote the product through these influential users by giving them incentives. We next proposed the concept of reinforced marketing to perform effective WoM marketing where multiple influential users collectively market a product. We also analyzed the important characteristics of influential users such as these users post most of the content of the group and able to influence most of the population of the group. We found that influential users posted over a wide range of topics and received a large number of user reactions. Finally, we determined the best time of the year to start marketing in OSGs to improve the effectiveness of marketing.



# Chapter 4

## Information Diffusion using the Best Time to Post

### 4.1 Introduction

A large amount of content is generated every day in social media. One of the main goals of a content creator is to diffuse information to a large audience and thereby receive a large number of audience reactions in the form of likes, comments, shares, etc. The main obstacle in getting high information diffusion is that a content has a very short lifetime, and within this short lifetime it has to compete with many other contents [8, 30]. If a content is created at the time when most of the target audience do not engage with the content, then the content will not diffuse to a large audience. In this chapter, we primarily look at the effect of posting time on information diffusion. We propose techniques to compute posting schedules that will lead to increase information diffusion. We use audience reaction as a measure to evaluate information diffusion.

In this chapter, we use publicly accessible Facebook pages to create our dataset. Facebook pages are maintained by brands, businesses, organizations, etc., to inform

customers about their products and services. There are primarily two types of social network relationships: *friend* relationship and *follower-following* relationship. Facebook pages use *follower-following* kind of relationship. Each page has admin(s) who create contents in the form of posts. Users can follow the page and create reactions in the form of likes, comments, and shares. We call these users as audience. Posted content is broadcasted to the news feed of followers, and it has to compete with many other contents to be at the top of followers' news feed.

In social media, most of the audience reactions are received within the first few hours of posting [4]. As discussed in Chapter 1, if a content is posted at a time when audience are not online or not interested in interacting with the content, the content will not receive a large number of audience reactions. Facebook's News Feed algorithm [100] rewards a post if it is getting a large number of audience reactions by increasing the rank of the post. If the post appears at the top of the news feed of many users, it would get more audience reactions and thereby becomes more popular.

Apart from looking at the ideal posting time for individual pages, it would be interesting to characterize the pages into groups with similar audience reaction profile. This will enable us to understand the factors that determine audience reactions. Given there are millions of Facebook pages, creating page category and then computing the posting schedule for the whole category will give higher statistical confidence while comparing the similarity and differences between various pages. With this characterization, we can also determine the ideal posting schedule for a new page that does not have enough audience interaction history. Let us consider the following example task:

**Example:** Consider a set of traffic related Facebook pages, where each page contains information about traffic updates for a particular city. Following are some of the questions that we address in this chapter:

1. What is the best time in a day that one should post about traffic updates to

get maximum audience reactions?

2. Is there any difference in the audience reaction pattern over the week?
3. Are there typical periods during the year in which people tend to look more at traffic updates?
4. How audience reaction pattern of traffic pages compare with other types of Facebook pages?

The key contributions of this chapter are as follows:

- We analyze post-to-reaction behavior of Facebook pages. We show that 84% of the audience reactions are received within 24 hours after posting.
- We identify top features that affect audience reactions and use these features to categorize pages into groups with similar audience reaction profile.
- We propose six posting schedules for individual pages and groups of similar pages.
- Our best posting schedule can lead to seven times more number of audience reactions compared to the average number of audience reactions that one would get without following any optimized posting schedule.

The remainder of this chapter is organized as follows. We formally define the problem of finding the right time to post to enhance information diffusion in Section 4.2. Section 4.3 presents the audience reaction behavior on Facebook pages. Section 4.4 discuss the categorization methods. Section 4.5 introduces the algorithm for schedule derivation. We proceed by describing schedule evaluations in Section 4.6. We finally conclude this chapter in Section 4.7.

## 4.2 Problem Formulation

In this section, we present the problem definition and details about the used dataset.

### 4.2.1 Problem Definition

The problem of finding the right time to post can be defined in terms of the following sequence of sub-problems:

**Problem 1 (Schedule for a Facebook page):** *Given a Facebook page  $P$ , find a set of time-interval(s)  $T_P$  such that if a post  $p \in P$  is posted during any time-interval  $t_k \in T_P$ , the post  $p$  is likely to get high diffusion, which is measured using the number of audience reactions received on  $p$ .*

Problem 1 is the right time to create a post for a single Facebook page. If a post is created according to the proposed schedule  $T_P$ , it would get more audience reactions. According to Facebook’s *News Feed* algorithm [100], if a post is getting a large number of audience reactions, the post will be given a chance to appear on top of the news feed of more number of users, thereby further increasing its likelihood to get high audience reactions. The schedule can be derived by using the posting behavior of page admins (pages) or the reaction behavior of audience. We state these two problems below.

**Problem 1.1 (Frequent Posting Schedule):** *Given a Facebook page  $P$  or a page category  $C$ , and the post creation profile  $M$ , find the frequent posting schedule  $S^{fp}$  for the page  $P$  or the category  $C$ .*

Admins of Facebook pages post a content at the time they receive the content (or just follow a certain personal schedule to post their contents). Although many admins may not be aware of when they should post to get maximum audience reactions, some expert admins with knowledge of social media post ranking might have an intuition of when they should post to get maximum audience reactions. They might realize

this by trying out various posting schedules. Thus, our first problem is based on the most frequent posting schedule (*category* is defined later in Problem 2).

Frequent posting schedule can be of three types: *aggregated*, *category specific* and *weighted category specific* denoted as  $S^{afp}$ ,  $S^{cfp}$ , and  $S^{wcfp}$  respectively. The aggregated schedule is the common schedule that can be used by all the pages. Categorized schedule is the customized schedule for categories, and it is the best schedule for all the pages in a given category. Within a given category, all the pages may not have the same importance. Weighted category specific schedule is derived by giving higher weight to more important pages within the category.

**Problem 1.2 (Frequent Reaction Schedule):** *Given a Facebook page  $P$  or a page category  $C$ , and the audience reaction profile  $R$ , find the frequent reaction schedule  $S^{fr}$  for the page  $P$  or the category  $C$ .*

Since our goal is to maximize the number of audience reactions, the frequent reaction based schedule is derived by analyzing the posting timings that lead to high audience reaction. Frequent reaction schedules are also of three types: *aggregated*, *category specific* and *weighted category specific* denoted as  $S^{ afr}$ ,  $S^{ cfr}$ , and  $S^{ w cfr}$  respectively.

**Problem 2 (Facebook page Categories):** *Given a set of Facebook pages  $\mathcal{P}$ , a set of reaction determining features  $F_R$ , categorize the pages in  $\mathcal{P}$  into  $r$  categories  $\{C_1, C_2, \dots, C_r\}$  such that similarity between reaction profile is high for pages within a category and low across categories.*

Each Facebook page has a unique pattern of audience reaction. The pattern is not the same for all the pages. Analyzing these reactions will help the page admins to get a deeper insight into their pages. For example, two e-commerce websites may have different type of audience reaction patterns, even though they may be from the same location or the similar type of organization. By categorizing pages into categories with similar audience reaction profile, we can understand what are the

different types of audience reaction profile? What are the factors that cause one page to get a certain type of audience reaction profile? If an organization wants its page to attain popularity similar to some other organization, what are the factors the organization should focus on to achieve that level of popularity? All these questions can be answered by looking at category-wise reaction behavior.

### 4.2.2 Dataset

We do our analysis on publicly accessible Facebook pages having a large number of audience. We obtain the dataset using the Facebook Graph API<sup>1</sup> in a similar way as described by Weaver et al. [101]. Each page has a profile page that contains posts created by page (posts created by the admin of page) and the reactions received on posts from the audience. Each page has a label (organization name) and a set of attributes (features). These attributes can vary across pages. A page can have attributes such as the number of fans (users who liked the page), the number of people talking about the page, type of the page, organization name, post creation time, reaction time, etc.

Audience can react on posts created by Facebook pages in the form of like, comment and share. Reactions consist of a textual comment and a unary rating score in the form of likes and shares. As an audience member reads a post, she can optionally create a reaction to the post created by a Facebook page. Each audience member can contribute one or multiple reactions to a post. Audience are allowed to update previous reactions and add new reactions on the reacted posts. Since we could only access timestamp for comments, we use comments as the reaction and the time of comments creation as the reaction timestamp. Comments can be used to implicitly measure the interest generated by a post [64, 102]. If we have access to all the reactions, we can incorporate them into our analysis using an aggregator function [93, 103].

---

<sup>1</sup><https://developers.facebook.com/docs/graph-api>

We extract the data of 100 Facebook pages from the same location that includes 5 different categories, namely e-commerce, traffic, telecommunication, hospital, and politician. Each of these categories contains the same number of pages to maintain homogeneity in audience reactions across the categories. Our collected dataset contains 0.3 million posts and 10 million reactions. As the dataset contains many unimportant and noisy words, we pre-process the data using text-processing techniques [96] such as stop-word removal, stemming, lemmatization, etc. We remove stop words from posts and comments as these words do not contain important significance to be used in the analysis. We also perform stemming and lemmatization to reduce inflected or derived words to their root forms.

## 4.3 Audience Reaction Analysis

In this section, we look at the user dynamics in Facebook pages. We analyze the time delay between when a post is created and when audience react to it. We also show different types of audience reaction that pages receive.

### 4.3.1 Post to Reaction Time Analysis

There is some time lag in post creation and audience reaction time [4, 8]. It is important to study this time delay as some of the important features used to find the right time to post are derived from this time delay. Typically, a post receives 97% of its total audience reactions within the first week of its posting. So, we consider timespan of one week to analyze post-to-reaction delay.

Figure 1.2 shows the distribution of audience reactions over a period of a week. We observe in Figure 1.2 that a post receives around 34% of its total reactions within the 1<sup>st</sup> hour of its posting, and 84% of reactions within a day. The lifetime of a post is very short, typically few hours and if it is not posted at the right time, it may not

get high audience reactions. So it becomes important for Facebook pages to choose a right time of the day to post a content. A Facebook page can post a limited number of posts per day/week. If a page creates fewer posts, it will not engage audience enough for them to maintain a social connection with the page and the page will lose engagement. On the other hand, if a page creates a lot of posts, it will typically lose engagement as audience can be overwhelmed with the page activities. So, it is important to know the right time (daily, weekly, monthly) to create a post in Facebook page. This is the motivation for our proposed problem to find the right time to post to enhance information diffusion.

### 4.3.2 Audience Reaction Behavior Analysis

We present audience reaction behavior profile of some real-world Facebook pages to understand the diversity of audience reaction pattern. We look at individual pages from politics, e-commerce, telecommunication, traffic, and hospital.

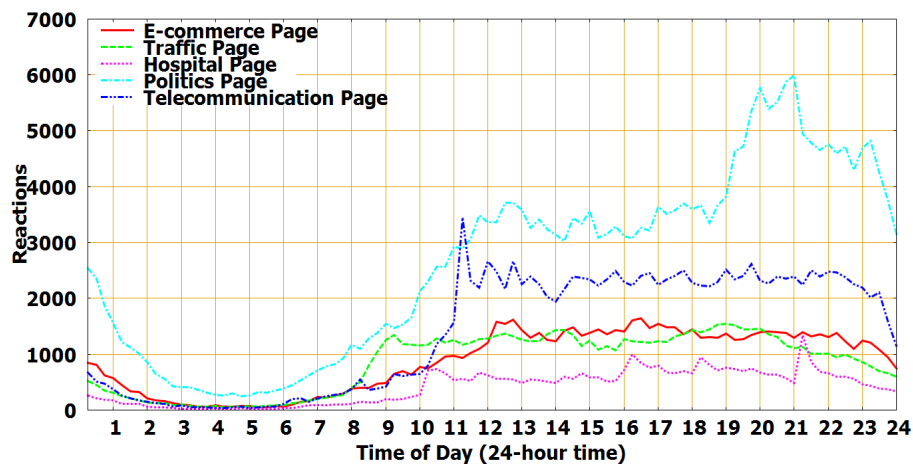


Figure 4.1: Audience reaction behavior

As can be seen in Figure 4.1, the audience reaction behavior vary across time and pages. Some pages have one or more peaks per day. Some pages have a uniform peak throughout the day. The page maintained by a politician, receive peak audience reaction between 8:00 pm - 10:00 pm. Audience reaction is much less during the rest



of the day. For e-commerce and telecommunication related pages the peak is around 11 am, and then it decreases a bit for the rest of the day. It indicates that the audience reactions also depend on the content and characteristics of the page [57–60]. We give more detailed results on audience reaction analysis in Evaluation Section 4.6.

## 4.4 Categorization of Pages

In this section, we give a solution to Problem 2. We present the reaction determining features and describe the method of feature processing, page categorization.

### 4.4.1 Reaction Determining Features

To find features that affect audience reactions, we create 35 features. We use a wrapper based feature selection to select the top reaction determining features. The features can be divided into following three types:

#### **Page centric features**

These are the features about the pages and signify popularity of the pages. Example features include the number of fans (those who have liked the page), the fan growth rate, the number of people who have created a story about the page on Facebook, and the number of posts per day.

#### **Content centric features**

These are the features about the page content. Example features include type of the page (described in Section 4.4.2); average number of likes, comments and shares for the whole page; average likes, comments and shares for different types of contents, such as Photos, Links, Videos, and average post length.

## Reaction centric features

These are the features about audience reaction. Example features include the average number of audience reactions received within various time intervals after the post is created, such as 0-1 hrs, 1-2 hrs, 2-4 hrs, 4-8 hrs, 8-16 hrs, and 16-32 hrs; the average number of audience reactions received during various day intervals, such as 12:00 am - 4:00 am, 4:00 am - 8:00 am, 8:00 am - 12:00 pm, and so on. These features also include the average number of reactions received on days of a week and months of a year.

### 4.4.2 Feature Pre-processing

We perform various pre-processing for the above features, such as correct the time zone, correct the type of page, convert continuous valued attributes to discrete valued attributes. We extract the timestamp associated with each post and reaction. Graph API provides the time in Greenwich Mean Time (GMT) format; we convert it into regional time-zone.

Admins of Facebook pages create the label (or type) for their pages, and they name it based on the domain of the page/organization. There are six primary labels provided by Facebook for pages namely, “Local Business or Place”, “Company Organization or Institution”, “Brand or Product”, “Artist, Band or Public Figure”, “Entertainment”, “Cause or Community”. Each of these labels includes multiple sub-labels such as “Brand or Product” includes “website”, “electronics”, “product/service”, etc. Each page admin has to select one of these labels for their page. There are inconsistencies between admins on how they select labels. For example, one e-commerce page is labeled as “Retail Company” and the other is labeled as “Website”. We use Nearest Neighbor algorithm [104] to label pages in a consistent manner, as page label is one of the most important determinants in our posting schedule analysis. We use topic modeling to represent the pages in terms of topics, and then use cosine similarity

of their topic probability to compute the similarity between pages. For each page, we find its  $k$ -nearest neighbor pages. We then use majority label from these  $k$  neighbors to correct the page label. If the page label is labeled correctly, then majority will also have the same label. If it is not the most appropriate label, then it will differ from the majority and we correct it by assigning the majority label. Since organizations from similar domain post similar type of information, this technique can give all the pages of same domain the most common label used in that domain.

In order to characterize the pages based on these feature attributes, we convert these continuous attributes to discrete attributes. We apply entropy-based data discretization [105] method to convert features in discrete attributes because most of the unsupervised data discretization methods require some parameter  $n$  such as the number of bins. Entropy-based method search through all possible values of  $n$  and capture inter-dependencies in features.

### 4.4.3 Categorization

We use clustering to group the pages with similar audience reaction. We use wrapper method [106] to select the features that are relevant for audience reaction. It considers selection of a subset of features as a search problem, where different combinations of features are used, evaluated and compared to other combinations. In the wrapper method, we use Multinomial Naive Bayes classifier [107] for classification. To create the base classes of Multinomial Naive Bayes, we use  $k$ -medoid clustering algorithm over the pages, where  $k$  is chosen using elbow method [108]. We use  $k$ -medoid algorithm instead of  $k$ -means algorithm because of its robustness to outliers as compared to  $k$ -means. Moreover, it uses representative objects as cluster centers instead of taking the mean value of the objects as a cluster center. Further, we define the similarity (refer to Equation 4.16) between two pages  $P_i$  and  $P_j$  in  $k$ -medoid as the similarity between their reaction profile  $R_k(P_i)$  and  $R_k(P_j)$  (reaction profile is defined in Section 4.5.1).

The top three obtained features are the *reaction within first one hour*, *number of posts posted by the page per day* and *type of the page* in increasing order of usefulness for the categorization. We cluster the pages using the top three reaction determining features as these three features are able to classify the pages into right category with the highest accuracy (90.3%) and increasing the number of features does not make a significant change in accuracy. The similarity in audience reaction within a category is high and across the categories is low when we use these three features for categorization (as shown in Section 4.6.3).

## 4.5 Schedule Derivation

In this section, we give a solution to Problem 1. First, we describe notations used in schedule derivation and later present six ways to compute posting schedule. The first two schedules are generic schedules that are applicable for all pages, whereas the last four schedules are category specific.

Let's assume we have data from  $d$  years  $\{Y_s, Y_s + 1, \dots, Y_s + d\}$ . We divide a day into 96 discrete buckets  $\{t_1, t_2, \dots, t_{96}\}$ , with each bucket of size 15 minutes as the bucket can capture essential reactions (as shown in Figure 1.2). By dividing a day time into small size of 96 buckets, we are able to determine right time (or bucket) more precisely. The first bucket  $t_1$  is from night 00:00 hrs to 00:15 hrs. We aggregate actions in the same time bucket from multiple years to ensure that our derived results are reliable.

We consider two types of actions: creation (posting) and reaction. We denote posting and reaction profile for a given time bucket  $t_k$ , page  $P_z$ , and year  $Y_j$  as  $m_k(P_z, Y_j)$  and  $r_k(P_z, Y_j)$  respectively.  $m_k(P_z, Y_j)$  is the aggregated number of posts created by page  $P_z$  at  $Y_j^{\text{th}}$  year (all days of year  $Y_j$ ) in the time bucket  $t_k$ . For each bucket  $t_k$ ,  $m_k(P_z, Y_j)$  is computed by counting the number of posts created by page

Symbol	Description
$P$	a given set of Facebook pages
$C_i$	a set of similar Facebook pages, $C_i \subseteq P$
$Y_s, d$	$Y_s$ is the base year in the dataset, $d$ is the total number of years
$t_k$	a time bucket of size 15 minute
$r_k(P_x, Y_j)$	reaction profile vector of page $P_x$ in $Y_j^{\text{th}}$ year
$R_k(P_x)$	cumulative reaction profile vector of page $P_x$ across $d$ years
$m_k(P_x, Y_j)$	posting profile vector of page $P_x$ in $Y_j^{\text{th}}$ year
$M_k(P_x)$	cumulative posting profile vector of a page $P_x$ across $d$ years
$\gamma^r(P_x)$	total number of reactions received in page $P_x$ in $d$ years
$\gamma^m(P_x)$	total number of posts created by page $P_x$ in $d$ years
$\rho^r(C_i)$	total number of reactions received in category $C_i$ in $d$ years
$\rho^m(C_i)$	total number of posts created by category $C_i$ in $d$ years
$W^m(P_x)$	fraction of posts created by page $P_x$ within its own category
$W^r(P_x)$	fraction of reactions received in page $P_x$ within its own category
$\delta(C_i, k)$	reaction per post for category $C_i$ in $k^{\text{th}}$ bucket across $d$ years
$\omega(C_i)$	aggregated reaction per post for category $C_i$ in all buckets across $d$ years

Table 4.1: Notations

$P_z$  in the time bucket  $t_k$  over the year  $Y_j$ .  $r_k(P_z, Y_j)$  is the aggregated number of reactions received in page  $P_z$  at  $Y_j^{\text{th}}$  year in the time bucket  $t_k$ . For each bucket  $t_k$ ,  $r_k(P_z, Y_j)$  is computed by adding all the reactions received by page  $P_z$  in the time bucket  $t_k$  over the year  $Y_j$ . We use these two profiles to compute the schedules.

### 4.5.1 Aggregated Schedules

We present two generic schedules that are common for all the pages. The first schedule ( $S_k^{afp}(P)$ ) is based on the aggregated frequent posting behavior and the second schedule ( $S_k^{afr}(P)$ ) is based on aggregated frequent reaction behavior of all the pages.

Aggregated frequent posting schedule ( $S_k^{afp}(P)$ ) is generated by using cumulative posting profile vector  $M_k(P_z)$ . For each time bucket  $t_k$ ,  $M_k(P_z)$  is the total number of posts created by page  $P_z$  in time bucket  $t_k$  across  $d$  years.  $M_k(P_z)$  is computed by aggregating the posting profile vector  $m_k(P_z, Y_j)$  of page  $P_z$  across  $d$  years as follows:

$$M_k(P_z) = \sum_{j=y_s}^{y_s+d} m_k(P_z, Y_j) \quad (4.1)$$

$S_k^{afp}(P)$  is a fraction of total number of posts created by all the pages in the  $t_k^{th}$  bucket. It is computed as follows:

$$S_k^{afp}(P) = \frac{\sum_{z=1}^N M_k(P_z)}{\sum_{z=1}^N \sum_{k=1}^{96} M_k(P_z)} \quad (4.2)$$

where  $P_z \in P$  and  $S_k^{afp}(P)$  is the fraction of total posts created by pages in  $k^{th}$  bucket, which is also defined as the probability of creating a post by pages in  $k^{th}$  bucket.

Similarly, aggregated frequent reaction schedule ( $S_k^{afr}(P)$ ) is generated by using cumulative reaction profile vector  $R_k(P_z)$ . For each time bucket  $t_k$ ,  $R_k(P_z)$  is the total number of reactions received by the page  $P_z$  in the time bucket  $t_k$  across  $d$  years.  $R_k(P_z)$  is computed by aggregating the reaction profile vector  $r_k(P_z, Y_j)$  of page  $P_z$  across  $d$  years as follows:

$$R_k(P_z) = \sum_{j=y_s}^{y_s+d} r_k(P_z, Y_j) \quad (4.3)$$

$S_k^{afr}(P)$  is a fraction of total number of reactions received by all the pages in the  $t_k^{th}$  bucket. It is computed as follows:

$$S_k^{afr}(P) = \frac{\sum_{z=1}^N R_k(P_z)}{\sum_{z=1}^N \sum_{k=1}^{96} R_k(P_z)} \quad (4.4)$$

where  $S_k^{afr}(P)$  is also defined as the probability of receiving audience reaction on pages in the  $k^{th}$  bucket. Now, we rank the buckets in decreasing order of  $S_k^{afr}(P)$ ,  $S_k^{afp}(P)$  with the first bucket being the best and the last one being the worst time to post according to these schedules respectively.

## 4.5.2 Categorized Schedules

As each category has different reaction behavior compared to other categories, we generate customized schedules for each category of Facebook pages. We derive two customized schedules for categories of Facebook pages, namely categorized frequent posting schedule and categorized frequent reaction schedule.

Categorized frequent posting schedule  $S_k^{cfp}(C_i)$  is computed based on number of posts created by category  $C_i$  in time bucket  $t_k$ , and total number of posts created by category  $C_i$  in all the buckets as follows:

$$S_k^{cfp}(C_i) = \frac{\sum_{x=1}^{|C_i|} M_k(P_x)}{\sum_{k=1}^{96} \sum_{x=1}^{|C_i|} M_k(P_x)} \quad (4.5)$$

where  $P_x \in C_i$ ,  $M_k(P_x)$  is the cumulative posting profile vector of page  $P_x$  and  $|C_i|$  is the total number of pages in category  $C_i$ .  $S_k^{cfp}(C_i)$  is the fraction of total posts posted by category  $C_i$  in  $k^{\text{th}}$  bucket, which is also defined as the probability of creating a post by category  $C_i$  in  $k^{\text{th}}$  bucket. Similarly, categorized frequent reaction schedule ( $S_k^{cfr}(C_i)$ ) is computed as follows:

$$S_k^{cfr}(C_i) = \frac{\sum_{x=1}^{|C_i|} R_k(P_x)}{\sum_{k=1}^{96} \sum_{x=1}^{|C_i|} R_k(P_x)} \quad (4.6)$$

where  $R_k(P_x)$  is the cumulative reaction profile vector of page  $P_x$ .  $S_k^{cfr}(C_i)$  is the fraction of total reactions received on category  $C_i$  at  $k^{\text{th}}$  bucket, which is also defined as the probability of receiving audience reaction on category  $C_i$  in  $k^{\text{th}}$  bucket.

We rank the buckets in decreasing order of  $S_k^{cfp}(C_i)$  and  $S_k^{cfr}(C_i)$ . We pick first few buckets from both the schedules, which are the right time to post for a category  $C_i$  according to these schedules. We compute categorized schedules for all the categories by following the same procedure. First time bucket of ranked schedules is the best time to post for category  $C_i$  to enhance information diffusion.

### 4.5.3 Weighted Categorized Schedules

We derive weighted categorized schedules by assigning weight to the pages of categories based on their importance. Some of the pages receive a large number of audience reactions and some of the pages post a large number of posts compared to other pages. To maintain homogeneity of actions and audience reactions across all pages in a category, we use a weight factor ( $W^r(P_x)$  or  $W^m(P_x)$ ) in computation of the schedules. Weight signifies the importance of each page in its category. It is computed by using two parameters  $\gamma$  and  $\rho$  as follows:

$$\gamma^r(P_x) = \sum_{k=1}^{96} R_k(P_x) \quad (4.7)$$

$$\rho^r(C_i) = \sum_{x=1}^{|C_i|} \gamma^r(P_x) \quad (4.8)$$

$$W^r(P_x) = \frac{\gamma^r(P_x)}{\rho^r(C_i)} \quad (4.9)$$

where  $\gamma^r(P_x)$  is the total number of reactions received by a page  $P_x$  and  $\rho^r(C_i)$  is the total number of reactions received by a category  $C_i$  (all the pages of the category). Similarly,  $\gamma^m(P_x)$ ,  $\rho^m(C_i)$ , and  $W^m(P_x)$  are computed using cumulative posting profile vector ( $M_k(P_x)$ ). Weighted categorized frequent posting schedule  $S_k^{wcfp}(C_i)$  for category ( $C_i$ ) is computed as follows:

$$S_k^{wcfp}(C_i) = \frac{\sum_{x=1}^{|C_i|} W^m(P_x) \times M_k(P_x)}{\rho^m(C_i)} \quad (4.10)$$

where  $S_k^{wcfp}(C_i)$  computes the probability of creating a post by a category  $C_i$  at the  $k^{\text{th}}$  bucket. Now, we compute weighted categorized frequent reaction schedule



$S_k^{wcfp}(C_i)$  for a category ( $C_i$ ) as follows:

$$S_k^{wcfp}(C_i) = \frac{\sum_{x=1}^{|C_i|} W^r(P_x) \times R_k(P_x)}{\rho^r(C_i)} \quad (4.11)$$

where  $S_k^{wcfp}(C_i)$  computes the probability of receiving audience reaction on category  $C_i$  in  $k^{\text{th}}$  bucket.

Weighted categorized schedules are similar to categorized schedules, the only difference is that weighted categorized schedules are computed by assigning a weight to each page of a category based on its importance in that category. We rank the buckets in decreasing order of  $S_k^{wcfp}(C_i)$  and  $S_k^{wcfp}(C_i)$  for all the categories. We pick first few buckets from both the schedules, which are the right time to post for a category  $C_i$  according to these schedules. We compute weighted categorized schedules for all the categories.

## 4.6 Evaluations

In this section, we evaluate our proposed schedules, page categorization method and present the audience reaction behavior over time. We also discuss how audience engagement varies with the type of post content.

### 4.6.1 Evaluation Metrics

We use *reaction gain* to evaluate the schedules and *correlation* to evaluate the quality of our categorization method.

#### Reaction Gain

Reaction gain metric is used to compute the performance of proposed schedules. It measures the change in reactions received in a particular time bucket, compared to the average reactions per post. Before computing the reaction gain for a schedule

( $S$ ), we first rank the time buckets of the schedule ( $S$ ) over a period of 24 hours and compute two parameters: *reaction per post* ( $\delta$ ) and *aggregated reaction per post* ( $\omega$ ). We next explain how to compute reaction gain for categorized schedules and later describe the computation of reaction gain for aggregated schedules. Reaction per post ( $\delta$ ) is the total number of reactions received on pages within category  $C_i$  at time bucket  $t_k$  in  $d$  years divided by the total number of posts created at time bucket  $t_k$  by category  $C_i$  in  $d$  years. For the  $k^{\text{th}}$  rank bucket as per schedule of category  $C_i$ , *reaction per post* ( $\delta$ ) is computed as follow:

$$\delta(C_i, k) = \frac{R_k(C_i)}{M_k(C_i)} \quad (4.12)$$

where  $R_k(C_i)$  and  $M_k(C_i)$  are the cumulative reaction profile vector and cumulative posting profile vector for the category  $C_i$  respectively.  $R_k(C_i)$  and  $M_k(C_i)$  are computed by aggregating the cumulative reaction profile vectors, cumulative posting profile vectors of all the pages in its own category respectively.

Aggregated reaction per post ( $\omega$ ) is the total number of reactions received on pages of category  $C_i$  divided by the total number of posts created by pages of category  $C_i$ .

$$\omega(C_i) = \frac{\sum_{k=1}^{96} R_k(C_i)}{\sum_{k=1}^{96} M_k(C_i)} \quad (4.13)$$

Now, we compute reaction gain (RG) for time bucket  $t_k$  and category  $C_i$  as follows:

$$RG(C_i, k) = \frac{\delta(C_i, k)}{\omega(C_i)} \quad (4.14)$$

where  $RG(C_i, k)$  signifies the increase or decrease in reactions received by category  $C_i$  when it posts in time bucket  $t_k$ , compared to the average reactions per post it receives.

Similarly, we compute the reaction gain ( $RG(P, k)$ ) for the aggregated schedules

by using  $\delta(P, k)$ ,  $\omega(P)$ ,  $R_k(P)$ , and  $M_k(P)$ .  $R_k(P)$  and  $M_k(P)$  are determined by aggregating the cumulative reaction profile vector and cumulative posting profile vector of all the pages respectively. Next, we compute average reaction gain for the categorized and weighted categorized schedules in time bucket  $t_k$  as follows:

$$RG_{avg}(k) = \frac{\sum_{i=1}^r RG(C_i, k)}{r} \quad (4.15)$$

where average reaction gain ( $RG_{avg}(k)$ ) for  $k^{\text{th}}$  time bucket is the average of  $RG(C_i, k)$  across all the  $r$  categories. We use  $RG_{avg}(k)$ ,  $RG(P, k)$  to evaluate the performance of categorized schedules and aggregated schedules respectively.

## Correlation

We use correlation metric to evaluate the effectiveness of the categorization method. We compute correlation across the categories by using the cumulative reaction profile vector of categories as follows:

$$Co(C_i, C_s) = \frac{\sum_{k=1}^{96} (R_k(C_i) - \bar{R}(C_i)) * (R_k(C_s) - \bar{R}(C_s))}{\sqrt{\sum_{k=1}^{96} (R_k(C_i) - \bar{R}(C_i))^2} * \sqrt{\sum_{k=1}^{96} (R_k(C_s) - \bar{R}(C_s))^2}} \quad (4.16)$$

where  $C_i$  and  $C_s$  are two different categories.  $R_k(C_i)$  is the cumulative reaction profile vector (audience reaction) of category  $C_i$  in  $k^{\text{th}}$  bucket and  $\bar{R}(C_i)$  is the average audience reaction of category  $C_i$ .

Similarly, we use the cumulative reaction profile vectors of categories of pages ( $R_k(P_x)$ ) to compute the correlation within the category. We determine the correlation within the category by taking the average of correlation computed between each pair of the pages that belong to the same category.

## 4.6.2 Effect of Schedule

We evaluate our proposed six schedules using reaction gain metric defined in Section 4.6.1. As there are no previous baselines on *best time to post* for Facebook pages, we consider the first two generic schedules, namely aggregated frequent posting schedule and aggregated frequent reaction schedule as baseline schedules. We compute average reaction gain for all the categorized schedules, aggregated schedules and pick top-30 time buckets.

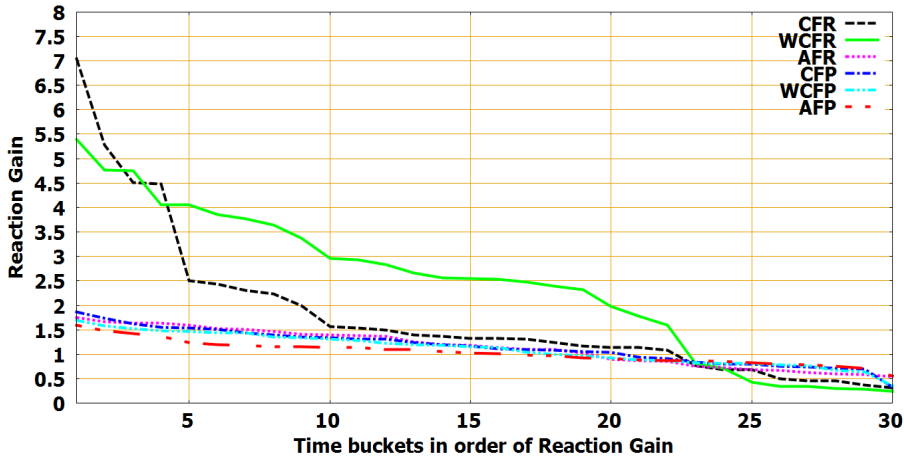


Figure 4.2: Reaction Gain

We observe in Figure 4.2 that all the posting based schedules, such as  $S^{AFP}$ ,  $S^{CFP}$ , and  $S^{WCFP}$  have reaction gain less than 2.0 even in their top bucket and their overall performance is also not as good as reaction based schedules. One of the reasons is that most page admins do not know what is the right time to post a content. They may not be even aware of the fact that they can get better audience reaction by just choosing a better time for posting.

On the other hand, reaction based schedules perform far better compared to posting based schedules. It is also observed that category-wise schedules perform better than aggregated schedules (baseline schedules). Reaction gain of categorized frequent reaction schedule ( $S^{CFR}$ ) is the highest (i.e., seven times better) in its top bucket. Weighted categorized frequent reaction ( $S^{WCFR}$ ) schedule shows a reaction gain of 5.4

in the top bucket.  $S^{WCFR}$  performs better than the  $S^{CFR}$  for all the buckets except the first two buckets. The reason could be that  $S^{CFR}$  is biased towards those buckets, which receive a large number of audience reactions. If a page or category receives a large number of audience reactions in a few buckets, it reflects high reaction gain in these buckets. However,  $S^{WCFR}$  is a normalized schedule, and it does not show high reaction gain if few buckets receive high audience reaction.

### 4.6.3 Effectiveness of Categorization

We compute the correlation within and across categories to show the effectiveness of our categorization method. Let us consider five categories: C1, C2, C3, C4, and C5. We label these categories using the type of most frequent pages in that category. With this, the categories C1, C2, C3, C4, C5 represent e-commerce, telecommunication, hospital, politics, traffic respectively. We consider two ways of doing categorization: using single feature and using multiple features. From the top reaction determining features, we select the best feature for single feature case. In multiple feature case, we consider all the top reaction determining features.

Categories	Single Feature	Multiple Features
C1 & C2	0.547	0.503
C1 & C3	0.392	0.341
C1 & C4	0.418	0.367
C1 & C5	0.519	0.470
C2 & C3	0.403	0.378
C2 & C4	0.353	0.302
C2 & C5	0.510	0.473
C3 & C4	0.351	0.305
C3 & C5	0.448	0.416
C4 & C5	0.440	0.419

Table 4.2: Correlation across the categories

Category	Single Feature	Multiple Features
C1	0.634	0.768
C2	0.703	0.848
C3	0.621	0.702
C4	0.650	0.771
C5	0.672	0.778

Table 4.3: Correlation within the category

We show across and within category correlation in Table 4.2 and 4.3 respectively

for both types of categorization. Ideally, we would want within category correlation high and across category correlation low. In the case of single feature, we find that within and across category correlation is almost same. However, in the case of multi-feature categorization, there is a large difference between within and across category correlation. These results indicate that our categorization function can categorize the pages effectively using multiple features. A new page that does not have enough reactions, can use this analysis to determine its right category and can post accordingly (as described in Section 4.6.4) to get a large number of audience reactions. For ease of presentation, in the rest of the chapter, we refer the categories as e-commerce, politics, etc. Each of these categories contains the same number of pages to maintain homogeneity in audience reaction across the categories.

#### **4.6.4 Trend Analysis**

We present some examples of audience reaction patterns, which are observed in daily, weekly and monthly analysis.

##### **Daily Analysis**

For daily analysis, we analyze the reaction behavior for all the above mentioned five categories, for 24 hours period over a duration of 5 years. Unlike Figure 4.1 that shows audience reaction behavior of individual pages, Figure 4.3 shows the aggregated audience reaction behavior of the categories. We observe in Figure 4.3 that categories can have audience reaction in different ways, such as multiple peaks, single peak and uniform peak during a day.

First, we analyze the categories that have multiple reaction peaks in a day (i.e., traffic, telecommunication). Reactions on traffic category are high during the start of office hours (11 AM) and end of office hours (6 PM to 8 PM). One of the reasons is that there is high traffic in these time-periods and people react in Facebook pages about

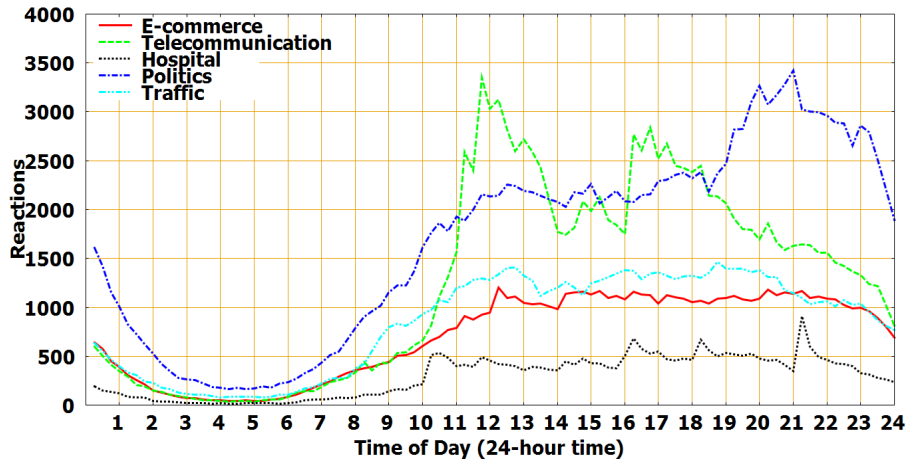


Figure 4.3: Audience reaction pattern on daily basis

the traffic problems that they have faced while going or coming back from offices. Similarly, telecommunications category has two peaks in a day: first is around 10 AM to 12 AM and second is around 4 PM to 6 PM. One of the reasons for this is that most of the people interact to social media pages in the morning to complain about an issue or to get the information related to tariffs, vouchers, special offers so that they can fill their balance and can use it throughout the day without out of balance problem. Some people prefer to do the same activity in the evening so that they can talk to family, friends, and relatives in the night when they become free from regular activities.

E-commerce category has uniform reactions from 12 PM to 10 PM (mostly during office hours) and drops after these hours. One of the possible reasons is that people usually take the opinion of their colleagues and friends working in the same office or organization about the product. If they found any issue, they often bring it to the notice of that e-commerce business immediately using Facebook page due to its quick response.

Pages related to politics and hospitals have single reaction peak per day. There is a high peak of audience reactions on politics category between 8 PM to 9 PM. One of the possible reasons is that people become free from their daily work by this time and

spend some time in knowing the political updates that are posted during the daytime. Similarly, people complain more about hospital-related issues in the evening, which they faced during the daytime.

## Weekly Analysis

In the weekly analysis, we analyze audience reaction behavior on two categories, namely telecommunication and traffic over the period of a week.

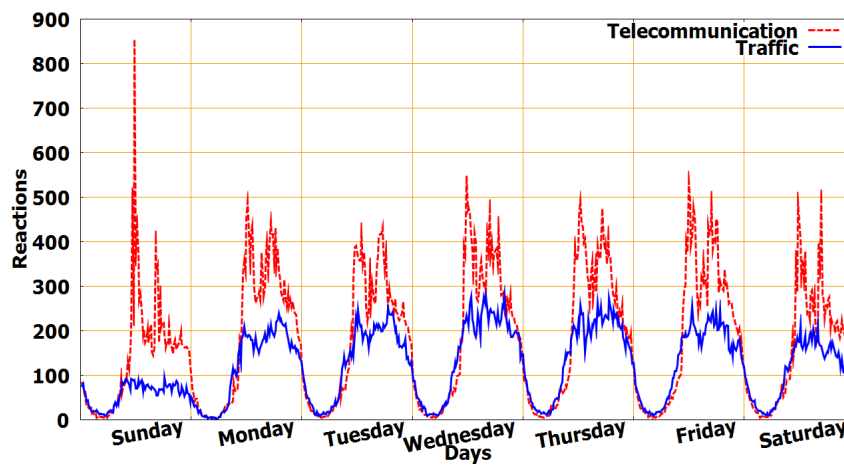


Figure 4.4: Audience reaction pattern on weekly basis

As can be seen in Figure 4.4, telecommunication category has the highest peak during Sundays compared to other days of the week. One of the reasons is that most of the people are free on Sundays and they prefer to fill their mobile and data balances. People react more to posts related to telecommunication such as special offers, vouchers during these days. Therefore, it is better to post important updates and offers on Sundays instead of other weekdays to get a large number of audience reactions.

Reactions on traffic category are high during working days and drop slightly during weekends. One of the reasons is that people do not go to offices on weekends as they have holidays. Audience reactions drop to half during Sundays compared to other days of the week because even on Saturday some people still go to offices, but most



of the people do not go to offices on Sunday. Most of the people stay at home and react less in traffic pages during weekends.

## Monthly Analysis

In the monthly analysis, we present the audience reaction pattern on two categories namely e-commerce and politics over the period of a year.

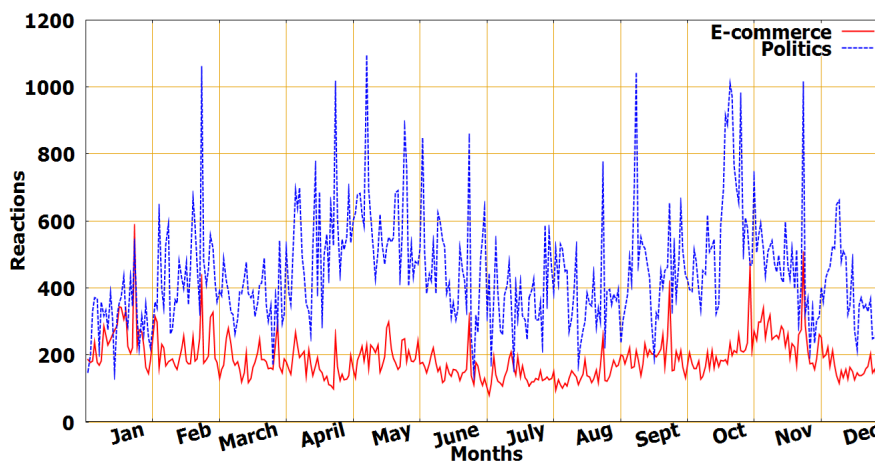


Figure 4.5: Audience reaction pattern on monthly basis

As it can be seen in Figure 4.5 that politics pages receive more number of reactions in the months of April and May. One of the possible reasons is that the politics pages included in dataset had their elections in these months. People are more active on social media pages during the election period. The peak in the months of October and November is due to the introduction of new fiscal policies. People react more about the advantages and disadvantages of new policies through the social media pages during these periods.

E-commerce category has more number of reactions during the months of October and November because these are the festive months in India and people buy new goods on the occasion of festivals. Increase in reactions during mid of December, January is due to Christmas and end of the year sale. These are the festive occasions and people like to purchase new items during these occasions. They would be interested

to know about offers and sales during these periods. If an e-commerce page post a news related to these sales and offers, people tend to react on it. Moreover, during these sales, lots of people purchase new items and a large fraction of these people face problems such as delivery issue, product issue, etc. People share their experiences<sup>2</sup> and complaints<sup>3</sup> about issues through social media pages of e-commerce.

#### 4.6.5 Audience Engagement with Contents

In this section, we show through empirical results that audience engagement depends on the type of content. Facebook page admins create different types of content, such as photo, video, link, and status. Some of these types of content receive more number of audience reactions compared to others. Pages can achieve higher audience engagement by creating contents of the type that receives more audience reaction. The results in Table 4.4 are based on the dataset mentioned in Section 4.2.2.

<b>Content type</b>	<b>Posts %</b>	<b>Reactions %</b>
Link	78.60	54.16
Photo	8.55	18.46
Status	1.59	1.21
Video	11.26	26.17

Table 4.4: Posts and reactions of different types of contents

In Table 4.4, for each content type, the second column shows the percentage of posts created by all the pages of that type, and the third column shows the percentage of reactions received by all the posts of that type. From the first row, we observe that although pages post 78.60% of the content as links, they get only 54.16% reactions from such content. In other words, links give less reaction (or audience engagement) per post. On the other hand, pages post only 8.55% and 11.26% content as images

<sup>2</sup>Today discounts looks impressive hope big billion days rock coming days?

<sup>3</sup>I ordered product exchange offer honor b 15oct 2015 but yesterday cancelled order without information.

and videos, which brings 18.46% and 26.17% audience reaction respectively. Videos can bring highest audience engagement.

## 4.7 Conclusion

In this chapter, we focused on enhancing information diffusion by posting a content at the time that increases the likelihood of getting high audience reactions. We analyzed user dynamics for individual Facebook pages as well as for a group of Facebook pages with similar reaction profile. We proposed six schedules for getting a high audience reaction, amongst which the best schedule leads to seven times higher reaction gain. We presented interesting audience reaction patterns in the form of daily, weekly and monthly temporal patterns. We also analyzed different types of contents to determine the content type that can increase audience engagement.

## Chapter 5

# Information Diffusion by Posting High Arousal Content

A popular and debatable content can widely diffuse the information. We say that a post has “high-arousal” content if it can attract a large number of user reactions especially in the form of comments. Arousal is similar to popularity with the difference being that arousal ensures lots of user feedback or comments. Among the three popular types of audience reaction in social media, namely likes, comments and shares, the reaction in the form of comments is the most informative as users can express their opinions in the form of comments. Users’ engagement can be increased by recommending them high arousal contents as high arousal contents are usually on debatable topic and ranked higher by post ranking algorithms (e.g., Facebook EdgeRank algorithm).

In the traditional print media, users do not have the option to express an opinion or read the opinion of other users. However, social media news channels allow users to express their opinions in the form of likes, shares, and comments. Likes and shares are positive reactions, where users most often agree with the news post. Opinion in the form of comments can be both positive and negative. Generally, comments are

more aggressive form of opinion compared to likes and shares. In social media news channels, users are not only interested in reading the news post, they are even more interested to read the comments of other users. In this chapter, we define arousal as a function of these three user opinions, where comment constitutes a large fraction of arousal. Given a news post based on its content we predict whether the post would generate high arousal or not.

Posts with high arousal are an asset to social media news channels and can give a big thrust to their businesses. A social media news channel can become more popular if it attracts a large number of users' opinions. When a news post starts getting many users' opinions, it attracts other users to read the post and give their opinions. We can enhance the existing news recommendation systems by recommending posts that have high-arousal and is of high interest to the reader. By improving news post recommendation, we can get more reader participation.

Apart from news post recommendation, there are many other advantages of showing high-arousal posts to users. This is the easiest way to get valuable user opinion, which can be mined to understand opinion of users. As e-commerce reviews are very useful in online purchase, the opinion of readers on news posts can help us to understand the opinion of the population on various important issues. Further, comments often give a more accurate picture of the reality compared to the original news post, which may be written in a biased manner due to various reasons. If a news post is genuine, a majority of the comments will support the post, otherwise they would disagree with the post. We therefore take up the task of predicting arousal for news posts published in social media.

The key contributions of this chapter are as follows:

- We define arousal of a news post in terms of social interactions such as likes, comments, and shares.
- We propose an unsupervised method to label the posts of high and low arousal.

- We present an ensemble-based classification method to predict the posts of high arousal with high accuracy.
- We determine the prominent topics using word-embeddings and named entity recognition, which can lead to high arousal for news posts.

The rest of the chapter is organized as follows. In Section 5.1, we present our methodologies. In Section 5.2, we proceed by describing the experimental evaluations and conclude the chapter in Section 5.3.

## 5.1 Methodology

Figure 5.1 shows the architectural overview of the proposed system. We perform the following steps to predict high arousal contents: (1) propose an unsupervised approach to create a training dataset by labeling high and low arousal posts; (2) generate the candidate features; (3) select features relevance to arousal prediction; (4) predict posts of high arousal; (5) determine the topics of high arousal from the posts of high arousal. In the rest of the chapter, we use terms such as ‘post’, ‘news post’, and ‘news content’ interchangeably.

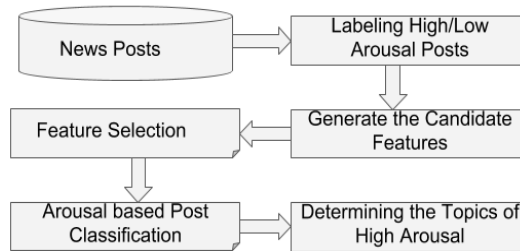


Figure 5.1: Overview of the arousal prediction

### 5.1.1 Labeling High and Low Arousal Posts

In this chapter, we use supervised learning to predict post arousal. However, there is no existing dataset that has labeled posts with high and low arousal. Moreover,

this kind of manually labeled training dataset is not so useful because the topics that arouse people today will not be the arousing topics after some time. A small fraction of news posts has content with high-arousal [109]. It is difficult to manually go through thousands of news posts and label them. Thus our first step is to use an unsupervised approach to label high and low arousal posts.

Arousal of news post is a highly subjective term. We define it in terms of popularity. There are many popularity measures such as *likes*, *comments*, *shares*, *clicks*, *views*, etc. Among all these popularity measures *likes*, *comments* and *shares* are publicly available measures. If a post has high popularity, it means that the post content is interesting enough that many people are interested in it. When many users are interested in the post content, it might also arouse other users to look at the content. Therefore, we use popularity as one of the deciding factors in the computation of arousal. We measure the popularity score ( $pt_{score}$ ) as follows:

$$pt_{score} = l + \sigma * c + \eta * s \quad (5.1)$$

where  $l$ ,  $c$ ,  $s$  are the number of *likes*, *comments*, and *shares* respectively.  $\sigma$ ,  $\eta$  are the comment and share popularity constants respectively. As suggested by Bucher et al. [93], *comment* and *share* require higher cognitive effort or commitment than *like*. Therefore, *comment* and *share* outweigh *like* suggesting that  $\sigma$ ,  $\eta$  values are greater than 1. Further, *share* generates higher amount of engagement compared to *comment* as shared post appears on user’s profile page. A shared post is pushed towards user’s connections as it constitutes a part of user’s self-presentation. This indicates that *share* outweigh *comment* (or  $\eta > \sigma$ ). In our experiment, we set the value of  $\sigma$ ,  $\eta$  to 2, 4 respectively which are derived from the analysis conducted by Kim et al. [103].

However, having high popularity does not ensure high arousal as there are many popular posts with a large number of *likes* and *shares*, but very few *comments*. *Likes* and *shares* signify that users like and agree with the post content. It is through

*comments* that the users express their agreement, disagreement or opinion with the news post. Therefore, out of the total number of reactions if *comments* constitute the major fraction, then we acknowledge the post as a comment dominant post. We compute the comment dominant score ( $cd_{score}$ ) as follows:

$$cd_{score} = c / (l + \eta * s) \quad (5.2)$$

A large value of  $cd_{score}$  indicates that post has received a considerable amount of user *comments*. From our empirical evaluation, we found that if a post is popular and it has a value of  $cd_{score}$  greater than 0.16, then it has a large number of *comments*. We found this number through a user study. We created different labeled datasets based on different values of  $cd_{score}$ . We asked 5 users to go through the different labeled dataset and find the parameter that gives the best training sample. To make things more tangible, consider real-life examples of news posts created by one of the most popular news channel CNN:

*Time famine refers to the universal feeling of having too much to do but not enough time to deal with those demands.*

The above post received 342 *likes*, 33 *comments*, 60 *shares*. According to our arousal Equation 5.2, the post should receive 93 *comments* but it received only 33 *comments*. It indicates that this post is not a post of high arousal. Let us look into another example post:

*It's hard to be as unpopular as President Donald J. Trump when the economy is going so well.*

The above post has 8900 *likes*, 3554 *comments*, 897 *shares*. According to our arousal Equation 5.2, the post should receive 1998 *comments* and it received 3554 *comments*. It indicates that this post is a comment dominant post, which garners lots of users'



opinions in the form of *comments*. As we look into the content of both the posts, this makes intuitive sense because the content of the second post is strong enough to arouse the people to provide their opinions about the issue as it contains predominant topics. This analysis suggests that our Equation 5.2 can identify the posts having a large proportion of *comments*.

Further, if we use only Equation 5.2, we may get posts that have relatively a large number of *comments*, but they may not be much popular. In other words, posts containing a large proportion of the *comments* are obtained. However, their popularity score, as defined in Equation 5.1 is low. Such posts will be of interest to only a small audience. In Equation 5.3, we define arousal score ( $a_{score}$ ), where we exalt the comment dominant score by multiplying it with the log of popularity score.

$$a_{score} = cd_{score} * \log(pt_{score}) \quad (5.3)$$

$a_{score}$  ensures that post should have a large number of *comments* as well as it should be popular. Logarithmic value of popularity ensures that if a post receives too many reactions, arousal is not increased too much. We rank the posts based on arousal, with the first post being the post of highest arousal and the last one being the post of lowest arousal.

### 5.1.2 Generate the Candidate Features

We generate the following candidate features from high and low arousal news posts:

#### POS Tag based Features

We use part-of-speech (POS) tagging to select features that are of interest to readers and news channels. Bandari et al. [71] showed that mentioning well-known entities in the post can increase its likelihood of becoming a popular post. For instance, consider

the following news posted by CNN: “I don’t want to do that at all. Trump said: I just want what’s right.” This post received lots of user attention because it mentions a well-known entity, Mr. Trump. Interestingly, important entities like ‘Trump’ are noun features [110]. We therefore perform POS tagging and then select the nouns and noun phrases as one set of candidate features.

The POS Tagger assigns a part-of-speech tag to each word of the given text, such as noun, verb, adjective, etc. Nouns are represented by ‘NN’, adjectives are represented by ‘JJ’, etc. We use the Stanford POS tagger [111] to do the tagging, and then extract the words with ‘NN’, ‘NNS’, ‘NNP’, and ‘NNPS’ tags as features.

## **Frequency based Features**

Important issues or entities appear frequently in news posts. We use term frequency and inverse document frequency (TF-IDF) [112] to find unigram and bigram terms that are important news topics. TF.IDF score, which is often used in text mining, shows how important a word is to a document in a corpus. The importance increases proportional to the number of times the term appears in the document (TF), but it is offset by frequency of the word in the corpus (IDF).

In this problem, we found that just using TF score gives better result compared to using the product of TF and IDF score. IDF score scales some terms inappropriately. Although IDF is very useful for ranking documents in information retrieval, it is not so useful in our problem. IDF gives more importance to terms that are rare in the corpus. However, to get high audience engagement, we need terms that are popular among audience. Although we do not use IDF score explicitly, we are able to remove less important words with high term frequency by using POS tags and seeing the term’s relevance to the post domain through the use of word-embeddings, which is described below in Section 5.1.3.

### 5.1.3 Feature Selection

All features obtained in the previous step may not be relevant for arousal prediction. In this section, we explain the irrelevant feature pruning steps.

We first take a combination of all the features from the previous section and build a bag-of-words (BoW) feature set, which is a set of features in high-dimension. We remove all features with sparsity more than 0.99 as it helps in generalization of classification task and prevents overfitting. We then use semantic relevance to remove more irrelevant features. News posts are categorized into various categories, such as entertainment, sports, politics, classifieds, etc. Users are recommended news posts based on the categories that they are interested in. An entity or topic popular in the sports domain may not be popular to users who are interested in some other domain, say politics or entertainment as number of reactions depends on the post domain. Since the posts are ranked with respect to other posts in the same domain/category, we use semantic relevance to prune features that are not relevant to the category.

For example, let us consider the following post from politics domain: “Kellyanne Conway told CNN’s Anderson Cooper that Donald J. Trump remains unconvinced that any breaches were part of an attempt to push him into the White House”. Here, *Anderson Cooper* is a CNN journalist, who is one of the primary CNN anchor and author. Even though the term *Anderson Cooper* is a noun feature and it appears very frequently in news posts, it has no connection with politics. Thus for news posts in the politics category, we identify terms that are not semantically related to politics and remove them from the list of potential features.

We use Google’s Word2vec model [113] to measure semantic similarity. Word2vec model creates word-embeddings by generating vector space from the text corpus where each word in the corpus is assigned to a vector in the space. We use publicly available word-embeddings<sup>1</sup> that were obtained by training Word2vec model with 100 billion

---

<sup>1</sup><https://code.google.com/archive/p/word2vec/>

words from Google News, and each of these vectors has a dimensionality of 300. We then find the similarity between features and categories using Word2vec similarity function. We select only those features that have similarity greater than 0.1.

#### 5.1.4 Arousal based Post Classification

In this section, we describe the method of predicting high arousal posts using post-content.

Using Equation 5.3, we find arousal score of all the posts and then sort them in decreasing order of the arousal. We select the first  $k$  posts as posts of high arousal and last  $k$  posts as posts of low arousal based on arousal score (in this chapter, we set  $k$  to 5,000). We determine the features from these two classes of posts as discussed in Section 5.1.2 and Section 5.1.3. We then train a binary classifier using these features. Moreover, our classification algorithm suffers from a class imbalanced problem due to the different variety of posts assigned to the classes, which leads to generate an unequal number of features for the classes. This results in biased prediction and misleading accuracy. To make an accurate prediction, we apply SMOTE algorithm [114] which maintains class balanced training set for the classification.

Further, integrating multiple classification techniques usually produces more improved and accurate results than a single classification technique. Dietterich et al. [115] have also shown that combining multiple classifiers give better prediction compared to a single classifier. We therefore use ensemble-based Voting Classifier [116] that weigh several individual classifiers and combine them in order to get a classifier that outperforms every one of them. The Voting Classifier implements ‘hard’ and ‘soft’ voting.

In this problem, we use hard voting in ensemble-based Voting Classifier. In hard voting, the final class label is predicted as the class label that has been predicted most frequently by the classifiers. In other words, the predicted final class label for

a particular sample is the class label that is predicted by the majority of classifiers when they perform the classification task individually on the same sample. In our experiment, we use Random Forest, Decision Trees,  $k$ -Nearest Neighbours, and Extra Tree classifiers. As each classifier gives different accuracy while performing the classification on the same sample, we assign the different weights to these classifiers based on their prediction accuracy. A new post that does not have arousal is passed to the Voting Classifier, which predicts whether the post would achieve high arousal or not.

### 5.1.5 Determining the Topics of High Arousal

In this section, we use posts with high arousal to find the topics of high arousal. The topical entities present in a post affect arousal of the post [71]. For example, a post related to ‘Donald Trump’ is expected to generate higher arousal compared to posts on trivial topics.

In order to determine the topics of high arousal, we use posts with high arousal. We perform Named Entity Recognition (NER) [117] over the posts and then use term-frequency to get topical entities that appear frequently in the posts. If two entities are consecutive, we merge them to form a single entity, as these consecutive entities often refer to a single entity, such as Donald Trump, Hillary Clinton, Brad Pitt, etc.

We also remove those unigrams that are part of some bigrams and they refer to a single person, place or thing (refer to Algorithm 2). For example, ‘Obama’ and ‘Barack Obama’ both represent the former president of United States. We remove the redundant term ‘Obama’ from the list of the topical entities. We do not remove unigrams that appear significantly without its superset bigram, such as gold, Olympics, President, Night, etc., which are the parts of gold medal, Olympics gold, President Obama, Saturday Night respectively.

Furthermore, NER may give many entities that are not related to a domain.

We therefore apply Word2vec similarity module to select the domain-specific topics. Using Word2vec similarity module on BoW features of high arousal posts, we also select topics that are not entities but show high similarity with a domain such as vote, surgery, competition, etc.

## 5.2 Evaluations

In this section, we first give details of our dataset. We then compare the performance of the proposed method and present the topics of high arousal from various categories of news posts.

### 5.2.1 Experimental Setup

Although our proposed method can be used in all types of social media news channels, we perform our experiments on the Facebook page of a news channel. In this chapter, we use the Facebook page of CNN<sup>2</sup>, which is one of the most popular television-based news channels. We crawl the publicly accessible data from the page using Facebook Graph API [101].

Our dataset contains the following information: post, audience reaction namely likes, comments, and shares, link to the actual news article, and various page attributes such as organization name, post creation time, reaction time, etc. Audience react to posts through likes, comments, and shares. Each comment reactions has an opinion text with the count of likes for that comment. We consider all the posts from April 2012 to December 2016, which aggregates to 33,324 posts and 226.83 million reactions. News posts are classified into different categories such as politics, health, sports, entertainment, etc.

We do arousal prediction for each category separately, as the number of audi-

---

<sup>2</sup><https://www.facebook.com/cnn/>

ence reaction varies greatly with the category. Some categories, such as politics, sports, etc., get a disproportionately a large share of attention from users. To clean our dataset, we also perform basic text pre-processing such as stop-words removal, stemming and lemmatization [96].

## 5.2.2 Effectiveness of Methods

As mentioned in Section 5.1.4, we use a voting classifier, which is a blend of multiple classifiers. Our voting classifier uses Random Forest, Extra Tree, Decision Tree, and K-Nearest Neighbour (KNN) classifiers. We assign weights to these classifiers based on their performances in our classification task. Random Forest and Extra Tree give the best performance followed by Decision Tree, which performs better than KNN.

To evaluate the importance of different feature sets and the proposed feature selection technique using Word2vec, we compute the accuracy, precision, recall, and F1 score [118] for the following feature sets: (1) POS tagged features, (2) TF based features, (3) Intersection of POS tagged and TF features, and (4) Union of POS tagged and TF features. Both POS tag and TF generate a large number of features. By taking intersection we get a smaller feature set, which contains the common features of these two sets. By taking union of these features, we get a much bigger feature set. Our aim is to find which of these feature sets give the best classification accuracy. For each of these feature sets, we consider both with and without feature selection. For feature selection, we use Word2vec to prune features that are not relevant to the domain. In the following table, ‘w/o’ and ‘w/’ means ‘without’ and ‘with’ respectively.

As can be seen from Table 5.1, POS without feature selection has 76.8% classification accuracy. Applying feature selection with POS does not improve accuracy. Instead, its accuracy decreases when the number of features decreases. POS tagging generates noun features that can capture useful entities such as, Donald Trump, Hillary Clinton, Brad Pitt, etc. If we apply feature selection then some of the less

Feature Set	Accuracy %	Precision %	Recall %	F1 score %
POS w/o FS	76.8	58.4	54.5	56.4
POS w/ FS	75.2	56.2	54.0	55.1
TF w/o FS	75.7	58.9	56.2	57.5
TF w/ FS	79.0	66.3	60.9	63.4
{POS $\cap$ TF} w/o FS	68.4	47.6	48.2	47.9
{POS $\cap$ TF} w/ FS	67.3	54.8	57.9	56.3
{POS $\cup$ TF} w/o FS	77.0	60.2	54.3	57.0
{POS $\cup$ TF} w FS	81.0	64.5	56.1	60.0

Table 5.1: Performance evaluation of arousal prediction

popular entity names are removed during feature selection, as they show very less similarity with a post category while using Word2vec similarity function, which results in lower classification accuracy. On the other hand, for the TF based feature set the accuracy and F1 score increases by 3.3% and 5.9% respectively with the use of feature selection. Using term frequency, we get many terms that are frequent but has no relevance to the post category. We are able to prune such irrelevant terms using Word2vec.

Intersection of POS and TF feature sets (w/ or w/o feature selection), gives the lowest classification accuracy. One of the possible reasons for this is that the intersection of POS and TF feature sets results in less number of features. Many important features that are frequent but not noun or noun but not frequent, are ignored while performing the intersection of two feature sets. Applying feature selection does not improve accuracy. On the contrary, it decreases accuracy because it further reduces the number of features and ignores some of the frequent noun features.

It is interesting to note that taking the union of frequent and tagged feature sets without feature selection, improves the accuracy compared to the intersection of feature sets. The reason for this is that the union of both feature sets generate a sufficient number of frequent or tagged features which are able to capture the relevant entities. However, the best result (81%) comes when we apply feature selection on the



union of POS, TF feature sets. One of the reasons for this is that Word2vec method in feature selection generates word-embeddings that preserve the aspect of the word’s context, which is an effective means to capture semantic relevance.

Overall, among the four feature set, the feature set obtained using union of POS and TF features gives the best classification, with 81.0% accuracy and 60.0% F1 score, and the one using intersection gives the worst classification, with 67.3% accuracy and 56.3% F1 score with feature selection, and 68.4% accuracy and 47.9% F1 score without feature selection. The TF feature set with feature selection gives the highest F1 score of 63.4% compared to the second highest F1 score of 60% using the union of POS and TF features.

### 5.2.3 Analyzing the Topics of High Arousal

We present some potential topics from four categories namely Politics, Health, Entertainment, and Sport that can increase the chances of getting high arousal on a post.

Category	Topics of High Arousal
Politics	Barack Obama, Bernie, Bush, campaign, Clinton, debate, Donald Trump, election, Hillary Clinton, immigration, Marco Rubio, poll, president Obama, White House, Republican candidates, United States, vote
Sport	competition, quarterback, FIFA, first olympic, gold medal, grand slam, legend, Leo Messi, match, medal, Michael Phelps, NBA, NFL, Olympic gold, phelps, punishment, relay, rugby, Ryan, Serena, soccer, superyacht, swimmers, tennis, Wimbledon, world series
Health	autism, kids, blood, body, brain, cancer, children, conjoined twins, doctor, drug, Ebola, experts, hospital, life, listeria, lunch, measles, medical, outbreak, pain, prevent, recovery, risk, surgery, transformation, treatment, tumor, virus, world health
Entertainment	actor, Angelina, Baldwin, Brad Pitt, breaking news, comedian, emotional, family, fan, favorite, films, Grammys, Harry Potter, history, Jennifer, Katy Perry, Lady Gaga, legend, Miss universe, NBC, night live, Saturday night, series, singer, talent, tv show, weekend

Table 5.2: Topics of high arousal from different categories

As can be seen in Table 5.2, all the categories show very prominent topics which are quite attractive and engaging. In the politics category, there are the topics such as Donald Trump, Barack Obama, Republican candidates, immigration, attacks, debate, etc., which can arouse people to interact or comment on posts related to these topics. Similarly, the topics captured under the entertainment category such as Brad Pitt, Katy Perry, Jennifer, Lady Gaga, comedy, movie, music, etc., are highly trending topics that can garner huge user attention. These celebrities often have a huge eager fan base who actively respond to every news about them.

Sport is another one of the most followed categories. News related to sport often covers major sport events and the players involved in these events. Sport events such as FIFA world cup, Olympics, Wimbledon and Michael Phelps the Olympic swimming champion are some of the appealing topics that can attract the user attention. Health is a major concern throughout the globe. We can see that news on issues such as the Ebola outbreak, the life-threatening diseases, symptoms and precaution measures for various diseases are critical. These topics catch the user's attention easily. Rather than following some general posting behavior, news channels can use this insight to achieve high arousal which can be helpful in various tasks such as predicting real-world outcomes, acquiring useful insights into users' collective behavior, efficiently allocating resources to support a better event or disaster management, etc.

### **5.3 Conclusion**

In this chapter, we studied the problem of increasing information diffusion through posting high arousal content. We predicted the arousal of news posts in social media using social interactions. High arousal on a news content indicates that people are interested in interacting with the content by providing their reactions especially comments. We modeled our problem as a classification problem and predicted if a news

post can achieve high arousal. We extracted the arousal determining features from the content of news posts. Using the best feature set for classification, we achieved an overall accuracy of 81%. We further analyzed the features or topics of high arousal from predominant categories such as politics, sport, health, etc. Interestingly, we found that news posts related to some particular topics such as popular celebrity, event, or controversial topics achieved high arousal.

# Chapter 6

## Information Diffusion of News in OSNs using Sentiment Dynamics

### 6.1 Introduction

Television, radio and print media are the three primary types of news sources in the world. Due to the unique nature of each communication medium and the manner in which their audience consume the information, they show a significant difference in their news communication. Despite these differences, all these sources are commonly accused of using exaggerated headlines to garner attention [119] and for focusing on negative news [120]. Today, social media has emerged as a powerful platform for the consumption of news with 68% of U.S. adults reportedly getting their news from social media [121]. Therefore, traditional news channels have started generating and disseminating news through various social media platforms.

As a growing number of people consume, share and discuss news online, it is important to understand whether the lack of regulation inherent in social media is being exploited to spread more aggressive and negative news. We surprisingly notice that news is not necessarily negative across all the news channels. Rather, there

is considerable variation in the way news is presented by different news channels and is heavily dependent on the medium through which these channels traditionally disseminated news, namely radio, TV or print media. Print and radio media based channels post more positive news while TV based channels post more negative news. The difference is not only because of the difference in the type of stories covered but because of how the same news is presented by different types of channels. To make the problem more tangible, consider the following example:

**Example:** Table 6.1 shows how the *Dakota Access Pipeline* protest, a movement in the Northern United States to protect natural resources and spiritual sites, reported by CNN and The Economist are extreme in tone but with opposite polarity (we measure polarity on the scale of -5 to +5) despite being posted at the same time and referring to the same incident.

<b>Organization</b>	<b>News</b>	<b>Polarity</b>
CNN	We're at the Standing Rock Sioux Camp in North Dakota. Protesters here are fighting to block the Dakota Access Pipeline and have vowed to stand their ground — despite growing calls for them to leave camp and threats of prosecution from law enforcement. Any questions for CNN's Sara Sidner?	-5
The Economist	Whatever the final result of the huge, long-running protests by native Americans against the Dakota Access Pipeline, the demonstrations will surely be remembered as a landmark in relations between organised religion, Christianity in particular, and indigenous people	+3

Table 6.1: Polarity of a news post generated by two different types of channels

Unlike one-to-many communication structure of traditional media, social media facilitates many-to-many communication by allowing users to express an opinion about the news by liking, sharing or commenting on them. The number of likes, shares, and comments received by a news post are good objective measures of user engagement and provide insight into what type of news interests users. We use this information to understand to what extent the sentiment policy employed by news media have been successful in catching users' attention or enhancing information diffusion. We

show that a negative news post receives a higher number of comments and shares compared to a positive news post, which gets a higher number of likes. This finding is extremely interesting as it supports the popular *negativity bias* theory<sup>1</sup> for heavily weighted actions such as shares and comments which require greater involvement but does not support it for the relatively simpler actions such as likes.

Comments allow users to express their opinion regarding a news post. These opinion can be used for opinion mining to gather information on how users perceive the news, predict real-world outcomes, gain useful insight into users' collective behavior, etc. These mining tasks often involve aggregating the users' opinions from different news channels, which may potentially bias the result because users' opinion on a topic depends on many factors [9,123,124], such as the region where the news is published, the time when it is published, the type of information source that published the news, the sentiment with which the news is written, etc. In this chapter, we analyze how users' opinion depend on two of these factors: the sentiment of a news post and the type of news channel. We obtain interesting insights that can be used to correct the bias arising due to the wavering nature of users' comments. To the best of our knowledge, very few studies [125,126] have been devoted to the sentiment analysis of news posts, and none has arrived to study the role of information sources coupled with the sentiment of news on the users' perception of the news. To gain more insight into the factors affecting the sentiment of news, we also categorize the news based on the topic, time, and their significance.

Our work can be used to enhance existing news recommendation systems. Users have preferences of what type of news they like to read. It could be based on the news topic and also the sentiment with which the news is communicated. For example, if a user prefers to read more positive and inspiring news, it would be better to recommend to such user news channels and news topics that mostly have positive sentiment.

---

<sup>1</sup>A popular theory in social psychology that states that humans are more likely to focus on bad news [122].

Following are some of our major findings: Sentiment generated by social media channels of different types of information sources are different and most of the time, these news channels generate either positive or negative news. Surprisingly, print and radio based channels generate predominantly positive news on their social media pages disagreeing with the popular opinion [127,128] that news sites mostly post negative news to take advantage of the negativity bias. We also found that negative news were shared and commented on more often, but positive news were liked more often throwing light on how the negativity bias operates at a different level of user engagement. Additionally, we show that polarity of the comments is strongly related to the sentiment polarity of the news. As news become more negative, their comments also become more negative in tone and vice versa. News from TV based news channels prompt more negative reactions from news compared to print and radio based channels, suggesting that people react not only to the type of news but also to the source of the news.

The key contributions of this chapter are as follows:

- We analyze the sentiment of posts created by different types of news channels in their social media pages.
- We investigate users' reactions to news posts of varying sentiment from different types of news channels.
- We categorize news posts into different topics to gain insight into the sentiment of news posts created under these topics.
- We compare the sentiment of big news headlines with niche news to investigate how big headlines impact the sentiment of news posts.
- We explicate the relationship between the sentiment polarity of news posts and the polarity of comments in conjunction with the type of information source.

The rest of the chapter is organized as follows. In Section 6.2, we present our methodologies. Section 6.3 analyzes the polarity of different types of social media news channels. Section 6.4 describes the relationship between popularity and polarity of news posts. In Section 6.5, we analyze the users' opinion on different types of channels with varying post sentiments. We present temporal analysis in Section 6.6 and conclude the chapter in Section 6.7.

## 6.2 Methodology

In this section, we first give the details of the process of collecting the news from Facebook pages of news channels. We then describe the method employed to measure the sentiment polarity. Next, we present the method to categorize the news posts. In the rest of the chapter, we use the terms 'post', 'news post', and 'news content' interchangeably.

### 6.2.1 News Posts Collection

In order to characterize the news posted on social media, we collected news from Facebook pages of five major news media channels. Our choice of Facebook over other social networking platforms is based on the research carried out by Pew Research center [121], which shows that Facebook has the highest reach with 44% of adults in the US getting news on the platform. Further, we chose news sites with the highest valuation as calculated by Virtue's Social Page Evaluator<sup>2</sup>. In order to understand the differences between different media, we chose two each of television and print media based channels and one radio based channel. Dataset thus includes posts from the Facebook pages of CNN and Fox News which are television news channels, The Economist and The New York Times (or NYT), which are daily and weekly newspaper

---

<sup>2</sup><http://www.adamsherk.com/social-media/most-valuable-news-site-facebook-pages/>



publishers respectively, and NPR which is a public radio network.

We extracted the dataset from Facebook pages using Facebook Graph API [95]. The dataset contains news posts posted by the pages, reactions on the post, link to the original news article and attributes including the number of users who liked the page, organization name, post creation time, reaction time, etc. Users can react to posts created by pages in the form of like, comment, and share. Reactions consist of textual comments and rating score in the form of likes and shares. For each news channel, we present the number of posts, comments, likes, shares, and time interval in the collected news dataset as follows:

News Channels	Posts	Comments	Likes	Shares	Time Interval
CNN	33324	26582081	147310056	52936764	Dec 2016-April 2012
NPR	18266	4585776	56007054	18847580	Dec 2016-Nov 2013
Fox News	26525	83957661	443933576	143762565	Dec 2016-Jan 2014
The Economist	24272	1336956	20206137	6376644	Dec 2016-Dec 2014
NYT	47522	9226029	93891025	25616593	Dec 2016-April 2013

Table 6.2: Dataset statistics

From Table 6.2, we can infer that Fox News is the most popular news channel as it has the highest reaction per post ratio, whereas The Economist is the least popular news channel among all the news channels as it has the lowest reaction per post ratio. Here, the reaction is the popularity measure, which is the sum of likes, comments, and shares. We performed pre-processing to remove noisy words from the textual posts and comments. We removed stop-words such as *a*, *an*, *the*, etc., as these words do not contain significant information for our analysis. We also employed stemming and lemmatization [129] to reduce inflected or derived words to their root forms. For the rest of the chapter, we consider a common time frame from December 2014 to December 2016 for our analysis.

## 6.2.2 Sentiment Polarity Identification

In social media, users use informal language to present their textual contents, which differentiates social media texts from standard texts. For example, they contain emoticons, such as *:)*, *:(*, *:-)*, *|-o*, and acronyms, such as *LOL*, *smh*, *ty*, *wth*. Many users use slang words, and these words became a part of social media lexicon. For example, *meh*, *yep*, *giggly*, *nah* are a few commonly used slang words. Users also use multiple punctuation marks to emphasize certain words in a text sentence.

In order to tackle all the above-mentioned issues, we used Valence Aware Dictionary and sEntiment Reasoner (VADER) [130], which is a popular and widely used sentiment analyzer for social media texts [131, 132]. VADER is a lexicon and rule-based sentiment reasoner and is suitable for sentiment analysis of contents originating in social media. It uses a new gold standard sentiment lexicon with 7500 lexical features that are commonly used to express sentiment in social media text. It uses a rule-based method to measure the sentiment intensity<sup>3</sup>. It primarily uses five generalizable heuristics based on grammatical and syntactic cues to convey how contextual elements increment, decrement or negate the sentiment of a text.

VADER has been compared with 11 sentiment analysis tools/techniques, including SentiWordNet [133], SenticNet [134], and LIWC [135] for social media text. It is shown that VADER outperforms all of them. VADER provides a sentiment score in the range of -1 to +1, with -1 being extremely negative, +1 being extremely positive, and 0 being neutral. For the sake of interpretability, we converted these polarity scores to an integer between -5 to +5. The polarity scores inferred were as expected and a sample of the same can be seen in Table 6.3.

---

<sup>3</sup><https://github.com/cjhutto/vaderSentiment>

Score	Sample Post
+5	It's just an amazing thing to watch good old-fashioned regular human beings and a whole lot of love change the world seismically
+4	Follow the Queen's Diamond Jubilee celebrations with the latest photos, videos, facts and trivia. Tell us which part of the festivities you're most impressed with
+3	When you do something extraordinary, it's shown that you can inspire other people." #CNNHeroes
+2	The world's first permanent ice hotel has opened in Sweden, thanks to new solar-powered cooling technology
+1	Farmers in the Australian desert are growing 15,000 tons of tomatoes using seawater — and thousands of mirrors
0	In a tweet, President-elect Donald J. Trump says his businesses won't do any new deals while he's in office
-1	Between 2007 and 2014, 30% of African elephants disappeared
-2	Being exposed to the daily hassles of traffic can lead to higher chronic stress and higher blood pressure," according to a recent study conducted in Texas
-3	Are we on the verge of a second Cold War?
-4	Terror attacks have ripped apart small towns and big cities across the Middle East and Africa throughout 2016, and this weekend was no different
-5	A young newly wed couple died a horrible death at the hands of the bride's family

Table 6.3: Sentiment polarity of sample posts

### 6.2.3 News Posts Categorization

To get insight into posting behavior of news channels across categories, it is useful to categorize the news posts into multiple categories such as Sports, Entertainment, Politics, Science and Technology, etc. Unlike online news sites, news posted on social media channels is not categorized. In order to categorize these news posts, we used LDA [80], which is an unsupervised text characterization method. LDA is a probabilistic topic modeling algorithm which represents each document (in this case, a news post) as a mixture of various topics with definite probabilities ( $\theta$ ). A topic is comprised of words or terms. The terms that often occur together, are placed under the same topic with high probabilities ( $\phi$ ). Document-topic distribution ( $\theta$ ) and term-topic ( $\phi$ ) are computed using Gibbs sampling [136] as follows:

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha} \quad (6.1)$$

$$\phi_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad (6.2)$$

where  $T$ ,  $D$  and  $\alpha$  represent the number of topics, documents, and smoothing constant respectively.  $C_{dj}^{DT}$  is the number of times a term appears in document  $d$  that has been assigned to topic  $j$ .  $W$ ,  $T$  and  $\beta$  represent the number of terms, topics, and smoothing constant respectively.  $C_{ij}^{WT}$  is the number of occurrences of a word  $i$  that has been assigned to topic  $j$ . Gibbs sampling method integrates these two assignments and updates the topic assignment until convergence.

In order to make the topic modeling richer, we augmented the post message with its URL information obtained using Graph API. The augmented text from external documents using URL was later processed to remove invalid characters and corrected for spelling mistakes. To determine the ideal number of topics  $k$ , we performed 5-fold cross validation on perplexity at different values of  $k$ . We then computed the

rate of perplexity change (RPC) [137] on a 10% random sample. Perplexity is a statistical measure and often used to measure the performance of topic models [138]. Perplexity reflects the capacity of a model to generalize to test set or unseen posts. The point where the rate of perplexity no longer falls significantly with an increase in the number of topics is used as the ideal number of topics. In our experiment, we found the optimum value of  $k$  is 10 from where perplexity does not change significantly.

As studied by Chang et al. [139] that perplexity and human judgment are not well correlated, we evaluated our topics manually using precision [140]. We asked five research scholars having knowledge of topic modeling to judge the relevancy of topical words generated by the topic model. We asked research scholars to label each topical word as relevant or non-relevant to assigned topic by the topic model. Topics are labeled by researchers independently without influencing each other. Topics for which researchers did not agree on were discussed until a consensus was reached. We then computed precision as a fraction of generated topical words that are relevant to the assigned topic. We found that the topic model performed reasonably well with 80.3% precision. One of the reasons for this high precision is that the number of topics selected for LDA categorization is ideal for the Facebook news posts dataset. Moreover, the posts created by news channels in their social media pages are well framed unlike user-generated contents, such as comments, tweets, etc.

Further, we provided the label for each topic based on the most relevant terms that uniquely define the topic. Since each topic contains thousands of terms, we extracted top relevant words based on term-topic distribution. Relevance ( $r$ ) of a term  $i$  to topic  $j$  is computed as follows:

$$r_{ij} = \lambda \log(\phi_{ij}) + (1 - \lambda) \log\left(\frac{\phi_{ij}}{p_i}\right) \quad (6.3)$$

where  $\phi_{ij}$  is the term-topic probability and  $p_i$  is the empirical probability of the word in the corpus.  $\lambda$  is a weighting term, and we chose 0.6 as an optimal value for  $\lambda$ ,

based on the results of Sievert et al. [141]. We assigned each post or document to these labeled topics based on document topic probability ( $\theta_{dj}$ ). If document  $d$  shows the highest probability for topic  $j$ ,  $d$  is assigned to topic  $j$ .

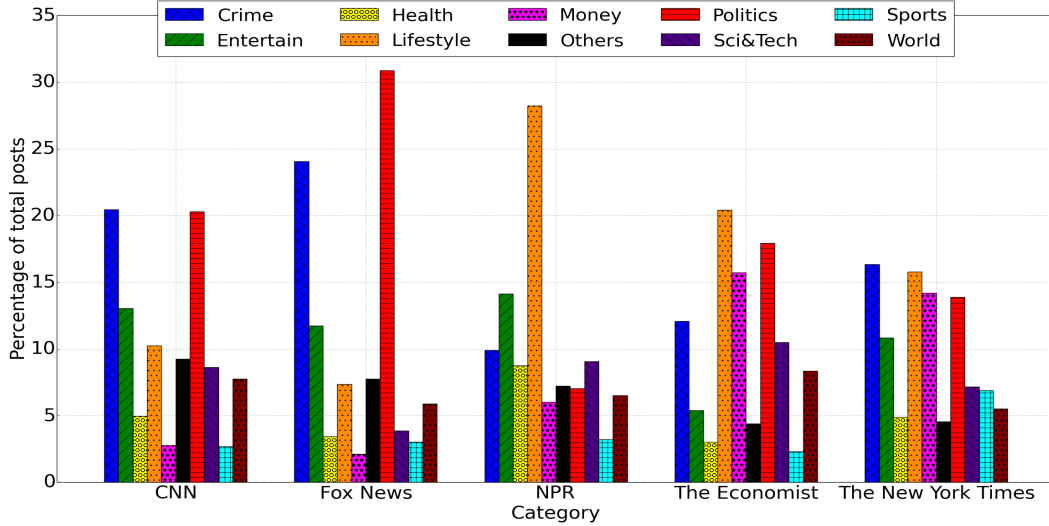


Figure 6.1: Distribution of news posts across categories

Figure 6.1 shows the distribution of posts for each channel across categories. We observe that post distribution of TV based channels CNN and Fox News is almost similar, where Politics and Crime categories contain a higher percentage of posts. In the case of print media based channels like The New York Times and The Economist, Lifestyle, Money, Politics, and Crime news seem to be more common. On the other hand, NPR which is a Radio based channel most often posts news related to Lifestyle, followed by Entertainment. We discuss how channels post news in these categories in Section 6.3.1.

### 6.3 Analysis of News Posts Polarity

We begin our investigation by analyzing the distribution of polarity of news posts grouped as positive, negative and neutral for five social media news channels as discussed in Section 6.1.

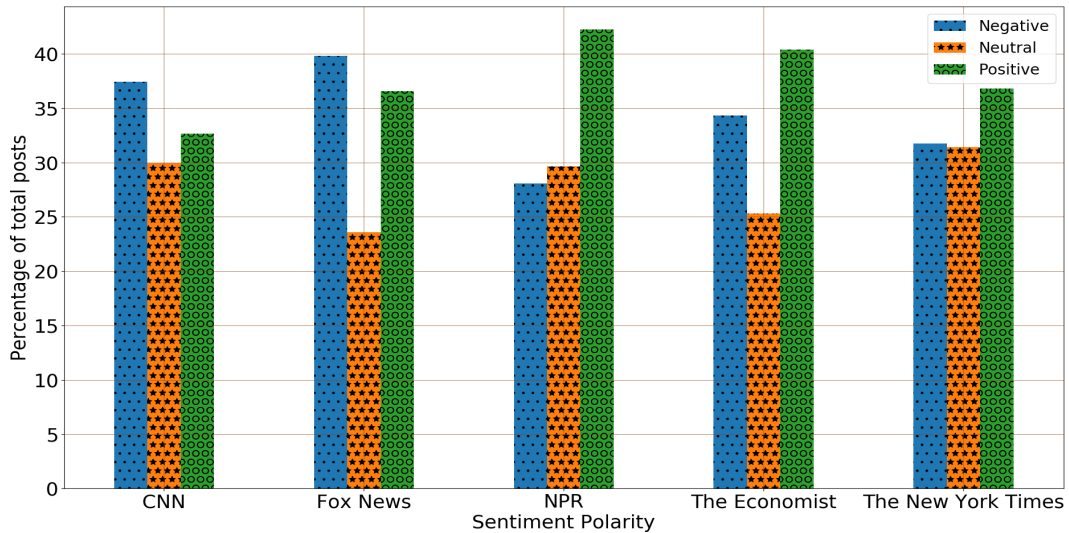


Figure 6.2: Polarity of news posts generated by pages

A quick glance at Figure 6.2 shows that the dominant sentiment of posts by all the news channels is always either positive or negative but not neutral. Moreover, posts with neutral sentiment are least common in Facebook pages of all media sites, except NPR. These inferences support the claim that all the channels tend to generate more positive or negative news on their Facebook pages to attract users' attention.

Another interesting aspect to be noted is the similarity in the distribution of sentiment polarity of posts between media channels that function through the same medium of communication. Posts by television based news channels, such as Fox News and CNN, are predominantly negative. Fox News generates the highest percentage (40%) of negative news across all the channels. On the other hand, posts by radio and print media based channels, such as NPR, The Economist and The New York Times are mostly positive. Radio based news channel, NPR generates the highest percentage (43%) of positive news and the least percentage of negative news (28%) across all the channels. Print based media channels, The Economist and The New York Times, generate a large proportion of positive news and a less proportion of neutral news. Despite the similar pattern of news generation by these two channels, The Economist generates a relatively higher percentage of both positive and negative

news compared to neutral news. One of the reasons for this is that The Economist reports growth (i.e., positive news) and decline (i.e., negative news) in commerce, and trade substantially.

The news reported by news sources has evolved differently because of the manner in which users consume information in each medium [142]. Our analysis suggests that these differences remain despite disseminating information on a common social media platform. News media are often criticized for focusing more on negative news rather than providing a balanced picture of the world [143–145]. This phenomenon has been attributed to journalistic cynicism and inherent preference for negative news among users. However, we observe through our analysis, that print and radio based social media channels post more positive news than negative news. This finding raises important questions: (a). Is this change in the type of content posted by print and radio based channels precipitated by user’s preference for positive news on social networking platforms? (b). Does this mean that the popular negativity bias theories [122, 146], which state that humans have a predilection for negativity, not hold true in the case of news consumption in social media? We attempt to answer these questions in Section 6.4.

### **6.3.1 News Posts Polarity across Categories**

In this section, we analyze the polarity of news posts across categories to investigate how news channels generate news across categories. We compare the polarity of news generated in multiple categories such as Sports, Politics, Health, Entertainment, etc. We observe that channels from the similar type of sources show the similar pattern. Due to brevity, we present the results of only one news channel from each type of information sources such as print, television, and radio.

It can be observed from Figures 6.3 and 6.4 that news belonging to Crime, World, and Health categories are predominantly negative, for both print and television based



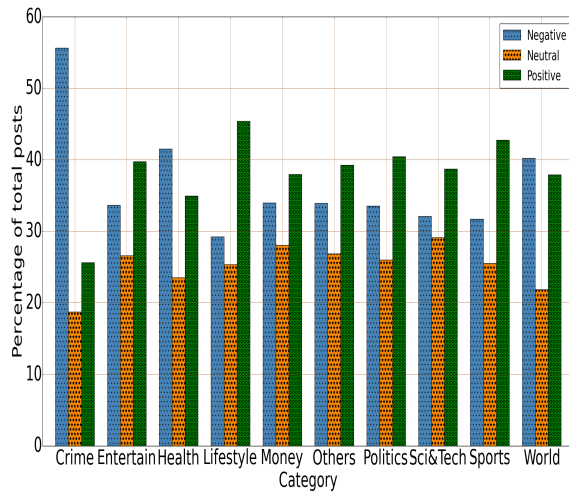


Figure 6.3: Category-wise post distribution of Fox News

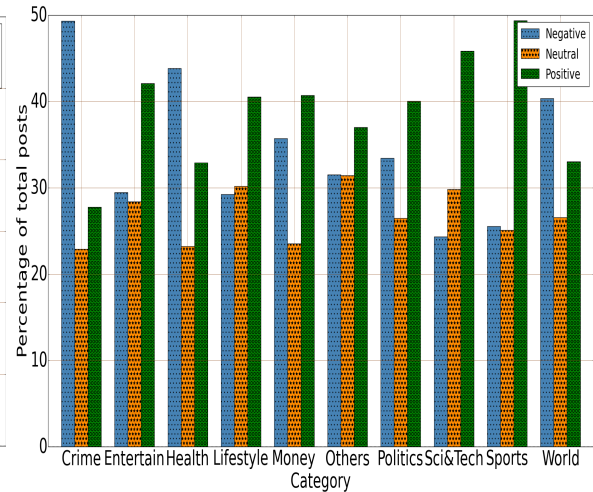


Figure 6.4: Category-wise post distribution of The Economist

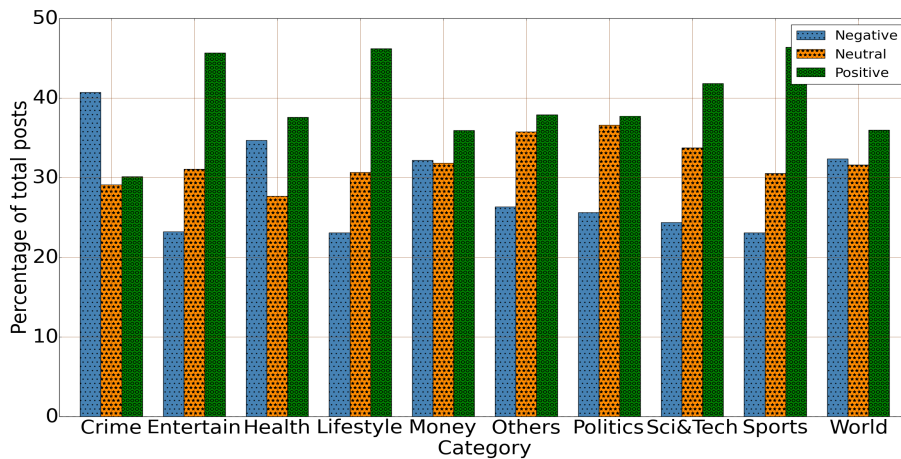


Figure 6.5: Category-wise post distribution of NPR

channels. One of the reasons for this is that most of the times news related to crime is woeful and unpleasant. News related to Health and World easily catch the attention of the channels if any negative event takes place across the world. However, in case of NPR (Fig. 6.5), which is a radio based channel, all types of news except Crime news, are predominantly positive in tone. This suggests the possibility that the trend of dominant sentiment observed in Section 6.3 could also be a result of the same type of news being covered differently by different news channels depending on the primary medium. Thus, by analyzing the sentiments across categories for different

news organizations, we can conclude that both the differences in the type of news that is often covered, and the difference in the tone with which news is covered are responsible for the difference in dominant sentiment observed in the previous section.

Moreover, except for NPR, across all categories, the proportion of news that is neutral in tone is the smallest. It indicates that the trend of a higher fraction of positive or negative news, which we observed in Section 6.3, is not limited to a few categories but is one of the tactics that is adapted for the generation of all types of news. NPR, however, stands out with negative news being the least common in majority of the categories. This observation is also consistent with the predominantly positive nature of news generated by NPR that we observed in the previous section. The similarity in the sentiment distribution for news channels using the same medium further asserts the influence of medium on the tone with which channels disseminate news.

### 6.3.2 Big Headlines Versus Niche News

In this section, we compare the polarity of big headlines and niche news. We compare the sentiment of news posts reported by different types of channels for big headlines as well as niche news.

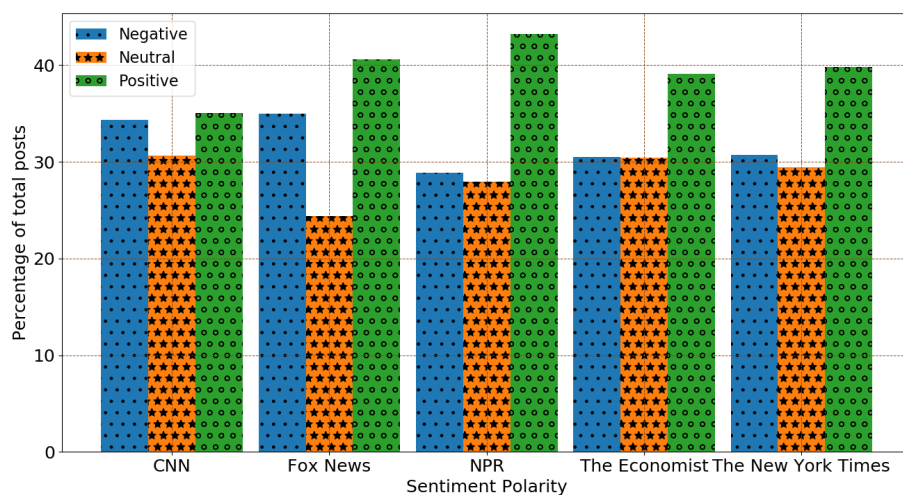


Figure 6.6: Big headlines

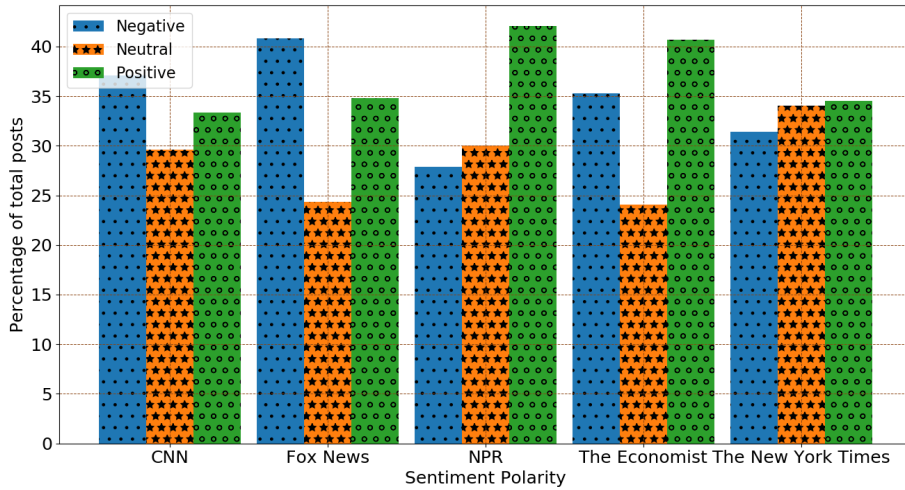


Figure 6.7: Niche news

As can be seen in Figures 6.6 and 6.7, the sentiment polarity of big headlines are different from the polarity of niche news. For big headlines, all the channels generate a higher percentage of positive headline news compared to negative and neutral news (see Fig. 6.6). Among all the channels, NPR generates the highest percentage (43%) of positive headline news. One of the reasons for generating a higher percentage of positive news is that these big headlines are very popular and exist for a longer time. If news channels continuously generate a higher fraction of negative news for these types of events, users may lose their interests and it would lead to less engagement on these news channels. Negative news have a shorter life time [70] and if there is a big headline that is usually persistent for some time, channels create a higher number of positive news to maintain the sustainability. As studied by researchers in psychology [147] that negative news causes worries to users, channels generate more positive news about the headline to retain the users' interests over time. On the other hand, we do not observe a significant difference in the polarity of niche news as compared to Figure 6.2. Positive or negative news is more popular than the neutral news. TV based channels report more negative news compared to the radio and print media based channels whereas print and radio based channels report more positive news.

### 6.3.3 Polarity of Same News Events across Channels

In this section, we examine how different types of news channels report the same news. We analyze sentiments of ten different real-world events that were posted by news channels in social media. To select the events, we used LDA followed by manual search. We used LDA to generate frequent topic phrases in our dataset. We then used these phrases to search for actual events in The Guardian API<sup>4</sup>.

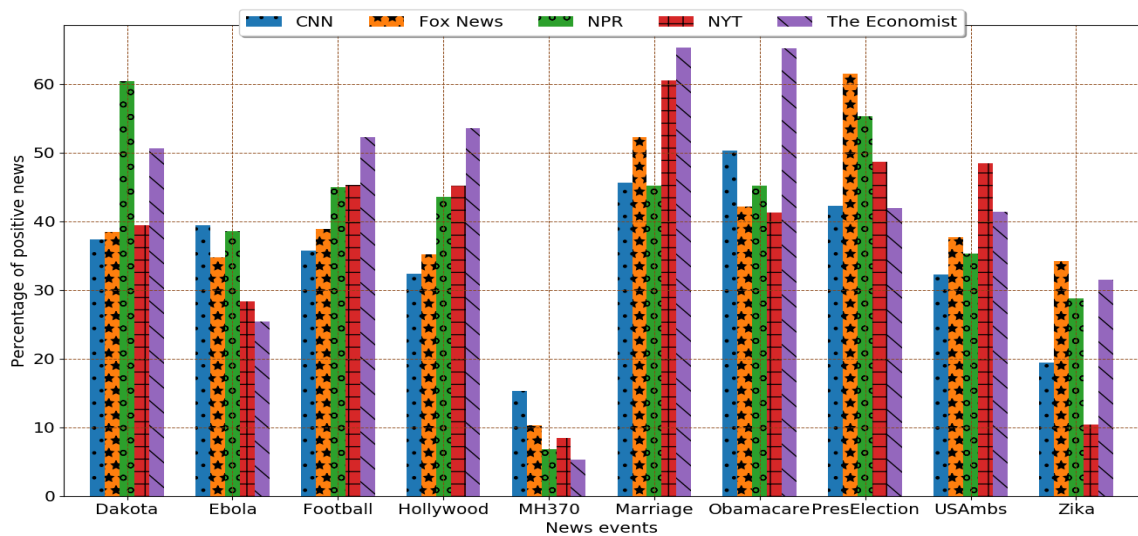


Figure 6.8: Percentage of positive news generated for a news event

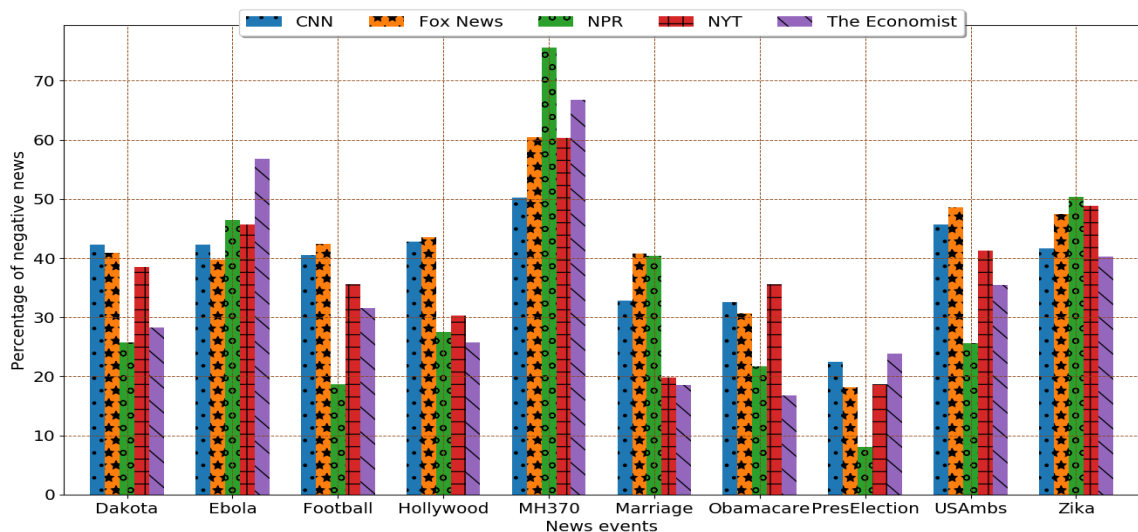


Figure 6.9: Percentage of negative news generated for a news event

<sup>4</sup><https://open-platform.theguardian.com/explore/>

Figures 6.8 and 6.9 show the fraction of positive and negative sentiment news generated by the five news channels for the ten events, where *PresElection*, *Marriage*, *Dakota*, *USAmb*, *MH370*, *Zika* indicate *Presidential Election*, *Same-sex Marriage*, *Dakota Access Pipeline*, *US Ambassador*, *MH370 Flight Disappearance*, *Zika Virus* respectively. For brevity, we don't show the fraction of neutral sentiment news, which can be obtained by subtracting the sum of positive and negative news from hundred.

Figures 6.8 and 6.9 show that different news channels report the same news events differently. Despite generating the same news with different percentage of sentiment polarities, all the channels generate a large fraction of positive news for big headlines such as *Presidential Election*, *Same-sex Marriage* and *Obamacare*. However, if a headline is very negative in nature such as *MH370 Flight Disappearance*, all the news channel generate a large fraction (more than 50%) of negative news due to nature of the news event. Also, all the channels generate mostly negative news for flu epidemics, such as *Ebola* and *Zika Virus* attack.

Although all the big-headlines are reported with a similar pattern of sentiment, regular events and the events that are not part of big-headlines are often reported differently by the channels. If a news post is related to regular events or minor headlines (e.g., *Football*, *US Ambassador*, *Hollywood*), channels usually generate different sentiment pattern of positive, negative and neutral news. In this case, either positive or negative news is more popular than the neutral news. TV based channels report more negative news, whereas print and radio based channels report more positive news.

Apart from the results shown in Figures 6.8 and 6.9, we also observed that news related to major breakthrough in Science and Technology, such as news related to *NASA* or *MIT*, highly reputed awards such as *Nobel Prize* or *Oscars*, and esteemed persons like *Pope* or *Dalai Lama*, are reported significantly positive (more than 55%) across all channels. One of the reasons is that these are very esteemed organizations,

awards, and persons. Due to their highly positive nature, all the channels generate mostly positive sentiment news related to them.

## 6.4 Popularity Versus Polarity

We analyze the popularity of a news post as a function of its polarity. Affinity metrics such as comments, likes, and shares received on a post are good indicators of its popularity. However, each of these actions involves a different level of interaction and are assigned different weights in the Facebook NewsFeed algorithm [103] with share receiving the highest weight and like the least. Hence, we do not aggregate these counts but analyze them separately. In order to account for the large difference in popularity of different news sites under consideration, we scale these counts of affinity metrics in the range of 0 to 1 and use the normalized values to determine the popularity of news posts.

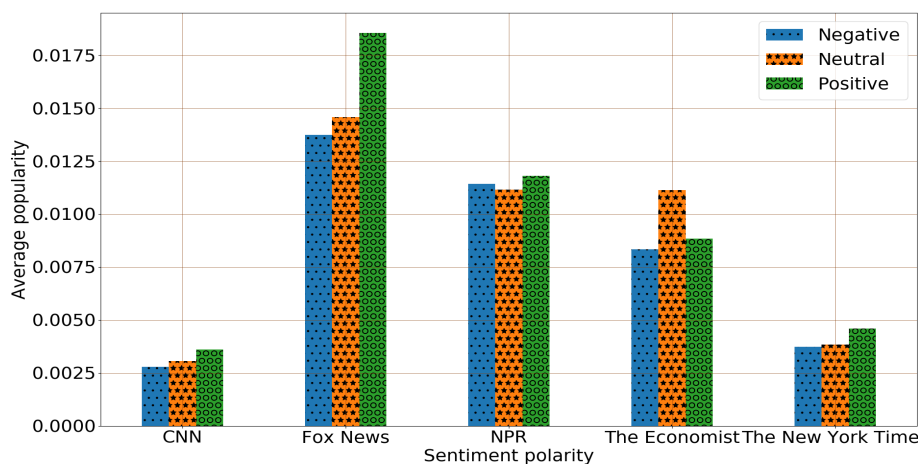


Figure 6.10: Likes on posts with different sentiments

We observe in Figures 6.10-6.12 that posts, which are either positive or negative, are more popular than the neutral ones in most of the cases. This suggests that news posts that are either positive or negative in tone tend to be more popular in social media. Exceptions to this are Fox News only for comments and The Economist

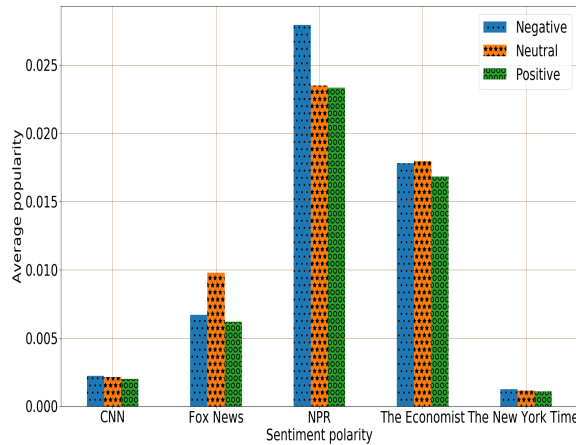


Figure 6.11: Comments on posts with different sentiments

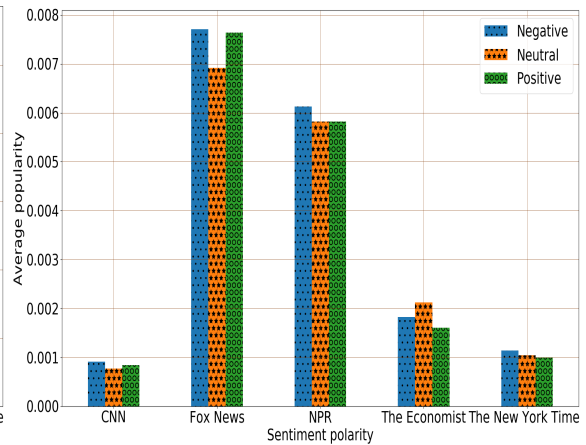


Figure 6.12: Shares on posts with different sentiments

channel. One of the reasons for this is that The Economist reported a large fraction of the news related to Money and Lifestyle (see Fig. 6.1). These business news are reported in the form of facts and figures (usually neutral in sentiment), which leads to a higher number of reactions for neutral news.

Further, in relation to preference between positive and negative content, we observe that more likes are received for positive posts whereas more comments and shares are received for negative posts. It shows that the results agree with *Negativity bias* theory for actions that involve a greater level of engagement such as commenting and sharing but disagree with the theory when it comes to simpler actions such as liking the post. Negativity bias is an elementary principle of psychology, which says that the psychological effects of negative news content outweigh those of positive news content [122, 148]. One of the reasons for disagreement of like popularity metric with negativity bias theory is that liking a negative news event shows that users have liked the negative news event. Comment and share reactions are different because creating a comment or sharing a negative news post shows that users want to express their opinions or spread the negative news post. Commenting or sharing a negative news post does not imply that the user has liked the negative news post. For example, if there is news about a plane crash or an earthquake, people may comment or share

the posts related to the news but it is highly unlikely that they would also like the same posts.

Trussler and Soroka [149] performed an eye tracking experiment to understand consumer demand for negative news frames and found a similar result. They found that participants said, they preferred good news but in reality often chose negative news stories over positive ones. While it is apparent that negative news receives a greater level of engagement, it is important to understand whether users are engaging positively or negatively with content in order to design any plan to receive appropriate users' opinions. We answer this question in the next section.

## 6.5 User Opinion Analysis

In addition to indicating the popularity of a post, user opinion (or comment) can provide a great deal of information about the tone of the audience, which can conclude whether the post is being perceived positively or negatively. To understand how users respond to posts of different sentiment polarities, we determine the average sentiment polarity of comments received for each post. Below we show the relationship between the average sentiment polarity of comments and the sentiment polarity of posts:

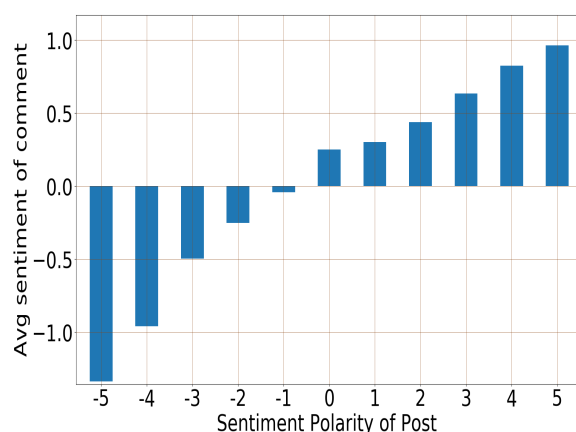


Figure 6.13: Avg. comment polarity vs. post polarity on CNN

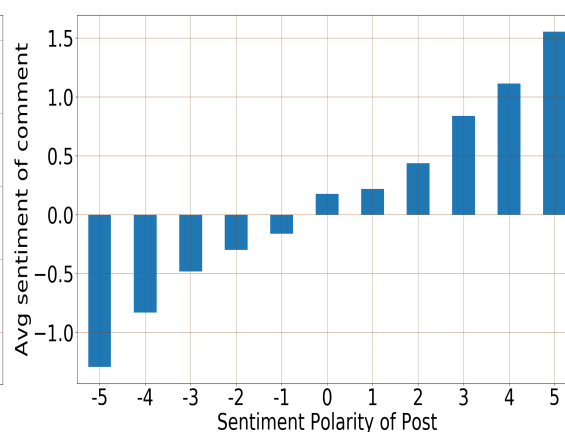


Figure 6.14: Avg. comment polarity vs. post polarity on Fox News



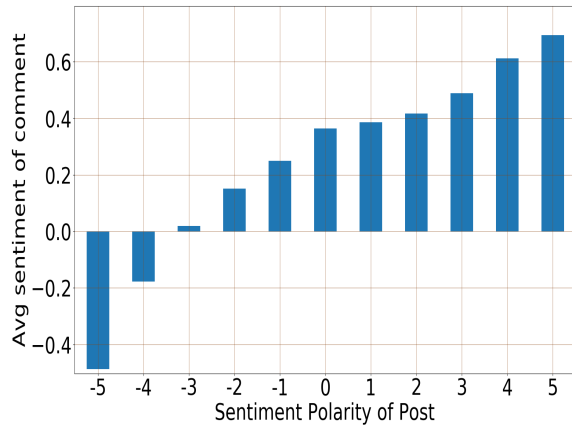


Figure 6.15: Avg. comment polarity vs. post polarity on The Economist

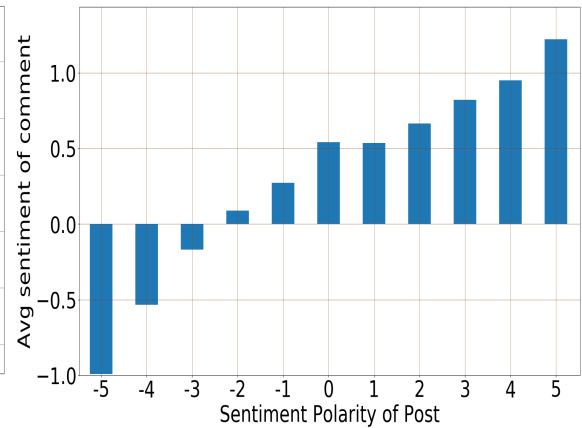


Figure 6.16: Avg. comment polarity vs. post polarity on The New York Times

From Figures 6.13-6.17, we observe that for all the five channels, as posts become more and more positive, comments also become increasingly positive. As can be seen in Table 6.4, there is a strong correlation between the sentiment polarity of comments and the sentiment polarity of posts. Comments of all the three types of channels have high sentiment correlation with the posts, and among all the channels TV based channels show the highest correlation.

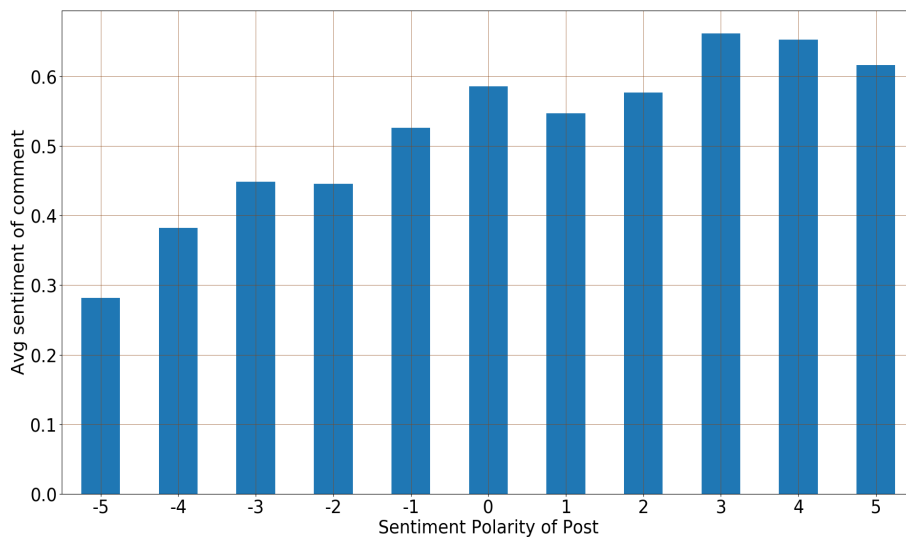


Figure 6.17: Avg. comment polarity vs. post polarity on NPR

We use statistical significance measure, namely p-test, to show the significance of our correlation result. A small p-value (i.e.,  $p < .05$ ) is sufficient evidence to

demonstrate the statistical significance of the results. The p-value for correlation coefficient is determined using the t-value as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (6.4)$$

where  $r$  is a correlation, and  $n$  is the number of observations (in our experiment, the value of  $n$  is 11). We use the Student t-distribution to find the p-value from the t-value, as suggested by Moore et al. [150]. We perform p-test on all the channels and observe that the p-value is much less than .05, which shows the sentiment correlation between posts and comments sentiments is significant at  $p < .05$ .

We can thus infer that posts written with varying levels of sentiment polarity prompt different reactions from users. A high correlation between post sentiment and comment sentiment suggests that measures of sentiment polarity of posts can be used to correct for biases that occur while aggregating comments from various channels for tasks such as opinion mining, opinion summarization, and real-world outcome prediction.

<b>News Channel</b>	<b>Correlation</b>
CNN	0.97
Fox News	0.98
The Economist	0.95
NYT	0.97
NPR	0.93

Table 6.4: Correlation between post sentiment and comment sentiment

However, the polarity for which a post starts attracting negative comments varies based on the medium of the channel. While Facebook pages of TV based channels, on an average, attract negative comments for negative posts and vice versa, comments for print media based channels do not become negative until posts become strongly negative in tone (i.e., sentiment score less than -3). It is interesting to note that the average comment sentiment polarity of NPR, which posts the highest proportion of

positive content amongst all the channels, remains positive irrespective of the polarity of post. It can be recalled from Section 6.3 that posts by TV based news channels were predominantly negative whereas those by print media and radio based channels were predominantly positive with radio based channel having the highest percentage of positive posts. This suggests that sentiment expressed in the comments is not only strongly influenced by the polarity of that particular post but also by users' opinion about the channel posting the news. The user opinion about the news channel is in turn shaped by whether the majority of the post messages have a positive or negative tilt. That is, Facebook pages of TV news channels which mostly post negative content attract more negative comments, whereas channels that are positive in tone like print media and radio attract fewer negative comments.

## 6.6 Temporal Analysis

In this section, we perform temporal analysis of news post polarity. We investigate how the polarity varies over the years. We also investigate whether posts of certain polarity drastically increase or decrease in particular months or days of the week. A common time frame from December 2014 to December 2016 is considered for the analysis.

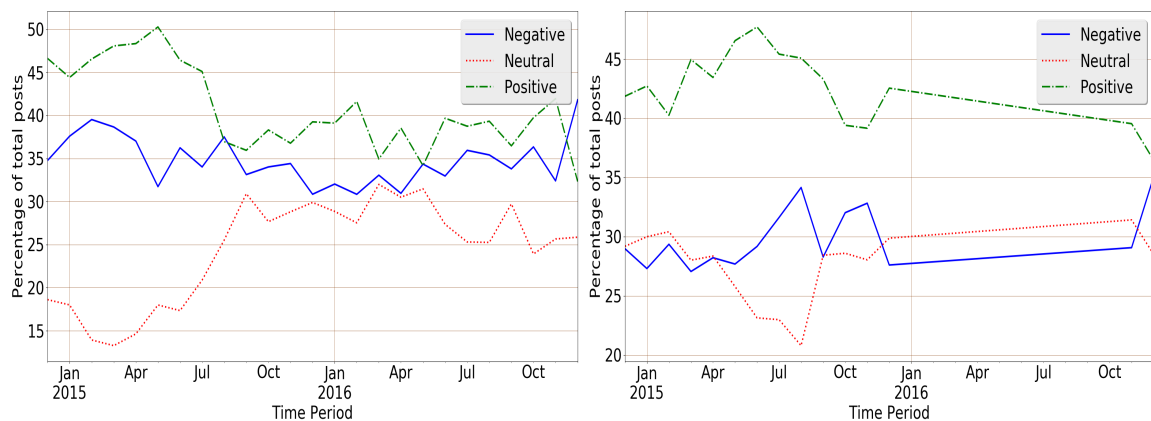


Figure 6.18: Temporal sentiment pattern of The Economist      Figure 6.19: Temporal sentiment pattern of NPR

Figures 6.18-6.21 show the polarity of news posts over the period of years. We observe that the behavior of television, print, and radio based channels remain the same as that was revealed in Section 6.3, i.e. negative sentiment dominates in the television based channels, positive sentiment dominates in the radio and print media based ones majority of the time.

We observe that, over time, Facebook pages of print based media channels (Fig. 6.18) show a gradual decrease in the percentage of positive posts while neutral posts increase and negative ones remain almost constant. As can also be seen in Figures 6.18 and 6.19, there is a peak of positive sentiment during the months of April to July 2015. One of the reasons for this is that two big headlines about *Obamacare* and *Same-sex Marriage* were in the trending news during that time. A few example news of these headlines are as follows: (a). The Supreme Court has ruled on Obamacare subsidies; (b). US supreme court declares same-sex marriage legal. These big news headlines lead to a sentiment peak in Figures 6.18 and 6.19, which is also in line with our analysis in Section 6.3.2 that news channels slightly generate more positive news for big headlines compared to negative news.

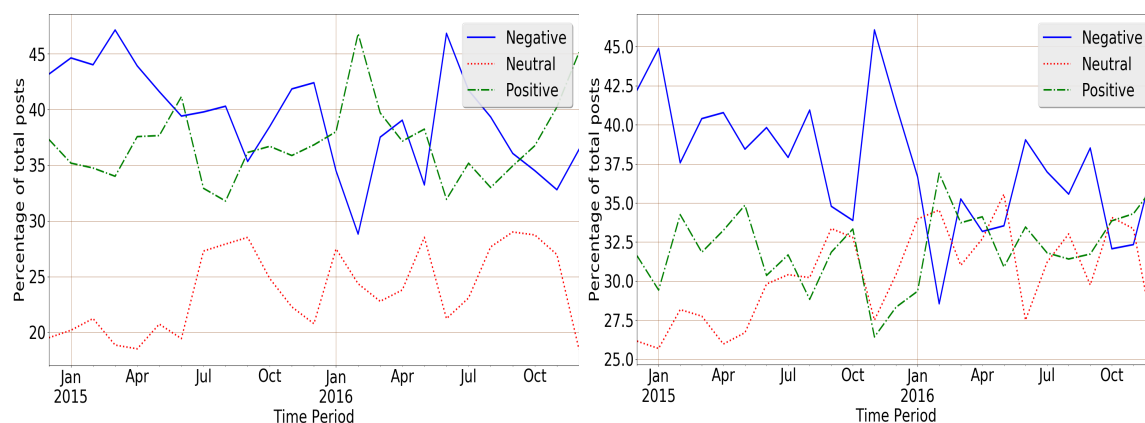


Figure 6.20: Temporal sentiment pattern of Fox News      Figure 6.21: Temporal sentiment pattern of CNN

Both the television based channels, CNN and Fox News exhibit a slight increase in positive news over negative news in the year 2016. Such instances of graphs of

same-medium news channels (Fig. 6.20 and 6.21) showing similar trends suggest that they behave and respond similarly to external events. One of the primary reasons for slightly dip in negative sentiment and a rise in positive sentiment is a big headline, namely *US presidential Election*. These TV based channels are somewhat biased towards a party and generate positive news about that party. Another big headline news was the *Nobel Peace Prize* award to the Colombian president. These big headlines seem to be one of the reasons for slightly dip in negative news and rise in positive sentiment.

We also analyze the news posts over the months and weeks, but we did not notice considerable variations in their polarities. On inspecting sentiment distribution over the months, we did not observe any consistent and significant change for certain months of the year. Even the weekly analysis does not reveal any significant change in the distribution, except for a slight increase in positive posts on weekends. One of the reasons for this is that news channels have posted slightly more Entertainment, Lifestyle, and Sports news during the weekend. This posting behavior of news channels leads to a slight increase in positive posts on weekends. Moreover, it is noted that the dominant sentiment in both the monthly and weekly analysis is also similar to the behavior of the channels observed in the previous Section 6.3. Thus, by analyzing the sentiment temporally, we can conclude that, on average, the specific characteristics observed in Section 6.3 are exhibited consistently across weeks, months and years.

## 6.7 Conclusion

In this chapter, we presented an extensive analysis on social media news channels from three types of news information sources to study the sentiment of news generated by these channels and its effect on users' reactions. We characterized news in different categories to uncover the distribution of the news posts and their sentiment polarity

across categories. Our analysis revealed that the sentiment of news posted by different types of social media channels is dependent on the medium through which these channels traditionally disseminated news. We also investigated the popularity of news with different polarity to get insight into the polarity of news that attract lots of users' attention. Interestingly, we found that news with positive or negative sentiment receive lots of users' attention. We also found that users' opinion depend on the sentiment of news posts and the type of information sources. Finally, we performed temporal analysis to understand how news posting behavior of social media channels evolve over a period of time.

# Chapter 7

## Information Summarization by Generating Topics of Interest

### 7.1 Introduction

Due to the huge size of social networks and a large amount of user-generated data, it is difficult to categorize, explore, and comprehend the data. Information summarization is a process of generating a concise and readable summary from a large amount of data such as social network data. We can summarize a large amount of data by presenting topical summaries such as key topics of interest. Topical summaries generated by existing methods are not readable and organized like manually created topics. In this chapter, we aim to generate topical summary similar to manually created summary from a large amount of bibliographic data about social network publications.

For example, in Figure 7.1, we show the topics of interest from the VLDB 2017 conference, which is one of the top conferences in Databases. However, these are manually created topics of interest. We aim to automatically generate such type of topics of interest. Although there are many popular topic modeling algorithms, such as LDA [80], Topical N-Gram(TNG) [81], Phrase-Discovering LDA [88], etc.,

- Access Methods, Concurrency Control, Recovery, Transactions, Indexing and Search, In-memory Data Management, Optimization, Storage Management.
- Privacy and Security in Data Management.
- Graph Data Management, Social Networks, Recommendation Systems.
- Data Mining and Analytics, Warehousing.
- Crowdsourcing, Embedded and Mobile Databases, Real-time Databases, Sensors and IoT, Stream Databases.
- Data Models and Query Languages, Schema Management and Design, Database Usability, User Interfaces and Visualization.
- Tuning, Benchmarking, Performance Measurement, Database Administration and Manageability.
- Distributed Database Systems, Cloud Data Management, NoSQL, Scalable Analytics, Distributed Transactions, Consistent Database-as-a-Service, Content Delivery Networks.
- Provenance and Workflows, Spatial, Temporal, and Multimedia Databases, Scientific and Medical Data Management, Data Cleaning, Information Filtering and Dissemination, Information Integration, Metadata Management, Data Discovery.
- Heterogeneous and Federated Database Systems.
- Fuzzy, Probabilistic and Approximate Databases, Information Retrieval, Text in Databases.

Figure 7.1: Topics of interest from VLDB 2017 conference website

all the conferences still use a manual approach to generate the topics of interest. There are many existing papers, such as [81,151], which have shown how one can use topic modeling to find topics of interest from the scientific literature or bibliographic data. Although the topics returned by these algorithms can be very useful for text categorization, they are not as readable and organized like manually created topics of interest. In this chapter, we present a novel algorithm based on frequent pattern mining and natural language processing to generate topics of interest that are very similar to manually created topics.

Topic modeling algorithms can be categorized into two broad categories namely, unigram-based topic models [80,152] and phrase-based topic models [81,87,88]. LDA is one of the most popular unigram-based topic models that was developed based on ‘bag-of-words’ assumption. For the papers published in VLDB, LDA algorithm generates topic words such as ‘based’, ‘methods’, ‘data’, etc. Many of these words do not convey complete information [153,154]. For example, the meaning of ‘graph data management’ cannot be completely captured by any one of the three words in isolation. Phrase-based topic models generate longer phrases but many of the top phrases, as explained later on, are ambiguous, redundant, and less understandable such as ‘case study’, ‘free data’, ‘large scale’, etc.

Once we have an algorithm to create human-like topics, we can use it in various



applications. For example, at present, conferences show a year-wise list of accepted papers, such as “VLDB 2018 accepted papers”. Similarly, one can show other useful information, such as, “key topics in VLDB 2018 accepted papers”, “key topics in VLDB 2010-2018 accepted papers”, “key topics from all Database conferences in 2010-2018 accepted papers”. We can use topic of interest information to compute various types of similarity, such as similarity between conferences, similarity between researchers and conferences, and similarity between researchers. All these above applications depend on our ability to generate topics in a principled manner, and these topics should match human generated topics, as the list of research topics of interest seen on homepages of researchers. In this chapter, we use association mining to generate a large number of possible topics, and then use NLP to refine and select the best topics.

Apart from generating topics from research areas and conferences, another problem that we address in this chapter is to group related topics. For example, in Figure 7.1, one can see that the topics Graph Data Management, Social Networks, and Recommendation Systems are shown as a group of related topics. All these three topics address graph-related problems. The existing topic modeling algorithms cannot be used to group such semantically related topics. In this chapter, we use word-embeddings to group semantically related topics.

As indicated above, the key contributions of this chapter are as follows:

- We show the limitation of existing probabilistic topic modeling (PTM) algorithms. These algorithms do not generate topics that are similar to manually generated topics.
- We present a novel topic modeling algorithm based on association mining and NLP. As compared to existing PTM algorithms, our algorithm with NLP refining generates topics that are almost twice more similar to manually created topics of interest and with 13.9% higher precision.

- We also present a novel algorithm to group similar topics using word-embeddings.
- We validate our hypothesis through experimental evaluations on a large DBLP bibliographic data and show a real-world application of proposed technique.

The remainder of this chapter is organized as follows. Section 7.2 presents the methodologies to find the topics of interest from research publications. We proceed by describing the experimental evaluations and the results in Section 7.3. Finally, we conclude our work in Section 7.4.

## 7.2 Methodology

In this section, we show the architectural overview of our system and present the various components involved in finding topics of interest using bibliographic research publications.

Figure 7.2 shows the architectural overview of our system. We perform the following five steps to determine the topics of interest: (1) categorize publications based on the conferences and research areas; (2) generate candidate topics for the individual conference or research area; (3) prune topics that are redundant, ambiguous, or uninteresting; (4) refine candidate topics to get more well-formed topics; (5) group semantically related research topics.

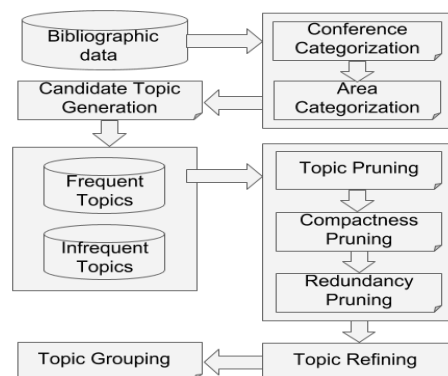


Figure 7.2: System architecture to find topics of interest

### 7.2.1 Data Categorization

In DBLP publications are not categorized (grouped) by conference or research area. Since our goal is to show topics based on conference and research area, our first step is to create two types of groups, one based on conference and the other based on research area. As DBLP data contains name of the conference, we can easily group papers based on the conference name. However, conference names are not given in a uniform manner. We therefore use standard entity resolution algorithms [155, 156] to clean our data and give a unique abbreviation to each conference (as can be seen in abbreviation data<sup>1</sup>).

To determine the topics of research areas, we categorize the publications based on their areas. We first find all the conferences related to an area and then assign all the publications of those conferences to that area. We observe that information about the area of publication is not available in DBLP dataset. Although a paper is published in one conference, the conference may belong to multiple research areas. For example, the conference KDD belongs to multiple research areas, such as Databases, Data Mining, Big Data (as can be seen in our area-wise categorized dataset<sup>2</sup>). The information about the research areas that a conference belongs to is not available on the web.

In this chapter, we use WikiCFP<sup>3</sup> to get the research areas for the conference. When a conference organizer posts *calls for papers* in WikiCFP, the organizers tag the conference with one or more of these areas (categories). We extract this information from WikiCFP. Even in WikiCFP, more than 50% of the conferences do not have any category assigned to them. We use collaborative filtering algorithms [157] from recommendation systems to find the possible areas for conferences that do not have any assigned category by using the category information from similar conferences with

---

<sup>1</sup><https://goo.gl/XdQGAW>

<sup>2</sup><https://goo.gl/sAAMnE>

<sup>3</sup><http://www.wikicfp.com/cfp/>

known categories. For a conference, WikiCFP shows the top-10 most similar conferences. The similarity is determined based on the fact that users who are interested in this conference are also interested in these other conferences.

## 7.2.2 Candidate Topic Generation

In this section, we first describe existing methods to find topics of interests and then we propose our method to find the topics.

There are two topic modeling methods namely, LDA and TNG, which have been widely used to find top topics from the text corpora. We use these techniques to find topics of interest of conferences and research areas from the publications. In topic modeling, a topic signifies the group of key-terms or phrases. We begin by describing LDA followed by TNG. We apply LDA on titles of the publications that belong to conferences and areas. As titles are more precise compared to abstracts, we use titles of the publications to find topics [158]. LDA represents each document (in this case, a publication) as a mixture of various topics with definite probabilities. The terms that often occur together are placed under the same topic with high probabilities. Document-topic distribution,  $\theta$  and term-topic distribution,  $\phi$  are computed using Gibbs sampling (refer to Equations 6.1 and 6.2).

However, LDA relies on the bag-of-words model and assumes that words are generated independently from each other. Therefore, it is not able to generate the meaningful phrases. For example, it generates words like ‘social’, ‘network’ as two different words even if they frequently appear together in the corpus. Topical n-gram (TNG) is a probabilistic model that determines topics containing words as well as meaningful phrases. TNG model is similar to LDA collocation model [159], the only difference is that TNG model can decide whether to form a bigram phrase for the same two consecutive words depending on their co-occurrences. As phrases convey more precise meaning compared to a single word, we use TNG model to generate meaningful

phrases. Although TNG generates better topics than LDA, we observe that it generates many phrases which are not so useful such as ‘based malware’, ‘case study’, ‘solving linear system’, etc. Therefore, we propose a novel method that generates better collocations by taking care of both co-occurrences and phrases.

Now, we describe our method of finding topics based on association mining [160] from the categorized DBLP dataset. We use association mining to find co-occurring words and phrases present in titles of the papers. Abstract or full-text contains lots of trivial words and phrases. If we use full-text of the papers, we end up getting several unimportant words and phrases. Researchers often use the main underlying concept precisely in the titles itself, so it is useful to apply association mining to find out these frequently occurring topic words or phrases [158]. The topics that are not frequent are likely to be rare or non-research topics.

We use the FP-Growth [161] association mining algorithm as it is one of the most efficient and scalable methods for mining frequent patterns [162, 163]. The algorithm generates frequent topics by using the criteria of minimum support and confidence. We perform experiments to determine the value of minimum support and confidence to generate substantial topics. We found that 0.1% minimum support and 60% confidence leads to a sufficient number of prominent topics. However, we observe that lots of distorted, redundant, ambiguous words or phrases are generated, and all of them are not interesting topics. Therefore, it is required to process these topics. We perform NLP based pruning and refining methods to get more useful and well-formed topics.

### 7.2.3 Topic Pruning

Not all the frequent topics generated by association mining are useful; there are topics that are ambiguous, uninteresting, and redundant. We prune these topics to prepare a list of selected adequate topics.

As mentioned in Section 7.1, single words often convey less information than phrases. For example, the meaning of ‘distributed database system’ cannot be completely captured by any one of these three words in isolation. We also observe that some common words frequently occur in the titles of the papers such as ‘approach’, ‘based’, ‘understanding’, ‘towards’, ‘method’, ‘challenge’, ‘opportunities’, ‘case’, ‘study’, ‘empirical’, etc. As a result, we get lots of uninteresting single words and phrases that contain these single words. Most of these words do not convey any useful topic information. Phrases contain a sequence of related words. Phrases are often more descriptive and carry a more precise meaning than single words. However, some candidate phrases are not interesting topics such as ‘mining frequent’, ‘component analysis’, ‘using support vector’, etc. Therefore, we perform compactness pruning and redundancy pruning to remove such uninteresting topics.

### **Compactness pruning**

Compactness pruning aims to remove those phrases whose words do not appear together in a specific order. It identifies the topics that contain at least two words and remove those that are meaningless or whose words are not in the right order. We perform compactness pruning to maintain the right word-order within a phrase as association mining does not capture the order of words in the publications. We first find the sets of phrases in such a way that in each set, phrases are equivalent but their word-orders are different such as {mining frequent, frequent mining}, {wireless sensor network, sensor network wireless}, etc. We then use term frequency-inverse document frequency (TF-IDF) to prune the disorder phrase(s) from these sets of phrases.

We consider all the publications (titles) of a conference or area as one document and generate the unigram, bigram and trigram phrases from each publication of the document. The reason for using unigram, bigram and trigram phrases rather than full titles or longer phrases is because most topics can be found based on local information.

Using long titles or phrases tend to generate a large number of spurious results. We compute the TF-IDF weight for each unique phrase, which signifies the importance of the phrase in a document. The importance increases proportionally to the number of times a phrase appears in the document but is offset by the frequency of the phrase in the corpus. We select top phrases from the TF-IDF generated phrases. We compare all of these phrases with previously generated sets of equivalent phrases. If any of the phrases in the set matches with the TF-IDF generated phrase, we consider that phrase as a potential topic and remove other phrases of the same set from the list of topics. In this step, we also perform an additional process of removing those topics which are common and do not convey any useful information such as ‘based’, ‘approach’, ‘case study’, etc. These topics have low IDF score because they appear frequently in the titles of all areas and conferences. To remove these kinds of topics, we find IDF score of all the topics. We prune the topics that have very low IDF score.

### **Redundancy pruning**

In this section, we explain redundancy pruning, which removes redundant words and phrases from the selected list of topics. We observe that topic list contains many bigram phrases which are part of the trigram phrases such as ‘component analysis’ is part of ‘principal component analysis’. We remove these types of insignificant redundant bigrams which are the part of the trigrams. However, we cannot remove all the bigrams which are part of the trigrams. For example, we cannot remove ‘neural network’ which is part of ‘artificial neural network’ as both are the significant phrases with distinct meanings. Therefore, we inspect the titles of the publications to know if bigram appears significantly in the titles without its superset trigram(s). We do not consider those bigrams that are subsets of trigrams and do not appear independently (independent of their superset trigrams present in topic list) at least  $th$  times (in our experiment, we set the value of  $th$  to 5) in the titles.

---

**Algorithm 2** Algorithm for Pruning the Redundant Topics

---

*Input:*  $T$ : Set of Topics  
 $TP$ : Titles of Publications  
*Output:*  $IrTopic$ : Irredundant Topics  
*Method:*

- 1:  $T_b \leftarrow T.GETBIGRAMS()$
- 2:  $T_t \leftarrow T.GETTRIGRAMS()$
- 3:  $IrTopic \leftarrow T$
- 4: **for all**  $t_b \in T_b$  **do**
- 5:      $st \leftarrow \{\}$
- 6:      $k \leftarrow 0$
- 7:     **for all**  $t_t \in T_t$  **do**
- 8:          $t_{tb} = GENERATEBIGRAMS(t_t)$
- 9:         **if**  $t_b \in t_{tb}$  **then**
- 10:              $st.ADD(t_t)$
- 11:         **end if**
- 12:     **end for**
- 13:     **for all**  $tp \in TP$  **do**
- 14:          $tp_t = GENERATETRIGRAMS(tp)$
- 15:          $tp_b = GENERATEBIGRAMS(tp)$
- 16:         **if**  $t_b \in tp_b$  and  $\forall_{a \in st} a \notin tp_t$  **then**
- 17:              $k=k+1$
- 18:         **end if**
- 19:     **end for**
- 20:     **if**  $k < th$  **then**
- 21:          $IrTopic.REMOVE(t_b)$
- 22:     **end if**
- 23: **end for**
- 24: **return**  $IrTopic$

---



Algorithm 2 shows the procedure to remove redundant bigram phrases from the set of candidate topics. Steps 1 and 2 extract bigrams ( $T_b$ ) and trigrams ( $T_t$ ) from the set of topics ( $T$ ) respectively. In step 3, we initialize the set of the irredundant topic ( $IrTopic$ ) by the set of candidate topics ( $T$ ). Steps 7-12 find the set of trigrams ( $st$ ) from the candidate topics that contain bigram. Steps 13-19 compute the number of times ( $k$ ) bigram appears in publication titles independent of previously determined trigrams ( $st$ ). If the value of  $k$  is less than the threshold ( $th$ ), we remove that bigram from the set of the irredundant topics (steps 20-22). Similarly, we prune unigram words from the list of topics that are part of some bigrams or trigrams topics and do not appear significantly in titles without its superset bigrams or trigrams.

#### 7.2.4 Topic Refining

Existing topic modeling algorithms or the algorithm described above will generate topics that are useful in applications where approximate or less understandable topics are sufficient. However, to generate topics that are similar to manually created topics, we present how to further refine the topics using NLP grammar rules.

We observe that some of the phrases appear in distorted forms such as ‘databases distributed’, ‘learning supervised’, ‘feature selection unsupervised’, ‘using support vector’, ‘programming genetic’, etc. We refine these kinds of topics to get cleaner topics using grammar rules from NLP. To apply grammar rules, we first need to tag the topics using Part-of-Speech (POS) tagger. POS Tagger assigns a part-of-speech tag to each word of the topic, such as noun, verb, adjective, etc. Parts-of-speech are represented by the tag such as nouns are represented by ‘NN’, adjectives are represented by ‘JJ’, etc. We use Stanford POS tagger [111] to do the tagging. We perform the following steps to refine the topics:

1. We observe that many phrases containing verbs appear in distorted forms such as ‘sensor networks distributed’, ‘data mining distributed’, ‘learning supervised’,

‘databases distributed’, ‘wireless network efficient’, etc. One common issue in all these phrases is that the verb does not appear at the beginning of phrases. We notice that if such verbs appear at the beginning of the phrase, it can form a better topic.

Among all the verbs, the verbs associated with ‘VBD’, ‘VBN’, ‘VBP’ tag frequently appear at the end of the phrases that lead to distorted topics. Therefore, we find the phrases that contain the last words associated with ‘VBD’, ‘VBN’, ‘VBP’ tag. These tags indicate that the corresponding word is either past form of the verb (‘VBD’, ‘VBN’) or non-third person singular present verb (‘VBP’). The last words of the phrases assigned with ‘VBD’, ‘VBN’, or ‘VBP’ tag are placed at the beginning of the phrases (move operation). We present the rules as follows:

$$w1\langle T \rangle w2\langle VBD \rangle \xrightarrow{\text{move}(w2)} w2\langle VBD \rangle w1\langle T \rangle$$

$$w1\langle T \rangle w2\langle T \rangle w3\langle VBD \rangle \xrightarrow{\text{move}(w3)} w3\langle VBD \rangle w1\langle T \rangle w2\langle T \rangle$$

We can see that each word is associated with a tag that is listed within  $\langle \rangle$ .  $\langle T \rangle$  shows that a tag ‘T’ is associated with a word and the word which is associated with ‘VBD’ tag is moved at the beginning of the phrase. By using this rule, we get the topics that are more interpretable than the actual candidate topics. For example, phrases like ‘data mining distributed’, ‘learning supervised’ are transformed into ‘distributed data mining’, ‘supervised learning’ respectively.

2. We refine the bigram phrases that contain progressive verb (gerund form of the verb) such as ‘programming genetic’, ‘scheduling sporadic’, ‘computing in-

ternational’, ‘learning active’, etc. These are the phrases whose first word is a progressive verb (‘VBG’) and the second word is an adjective (‘JJ’). We observe that the verb appears after the adjective leads to form a better topic. Therefore, we swap the first and second word of the phrase.

$$w1\langle VBG \rangle w2\langle JJ \rangle \xrightarrow{\text{swap}(w1,w2)} w2\langle JJ \rangle w1\langle VBG \rangle$$

For example, phrases like ‘scheduling sporadic’, ‘programming genetic’ are transformed into ‘sporadic scheduling’, ‘genetic programming’ respectively, which are well-formed phrases compared to actual candidate phrases.

3. We observe that there are many trigram phrases containing progressive verb appear frequently in the titles of the papers such as ‘solving linear system’, ‘mining association rules’, ‘training neural network’, ‘neural network modeling’, etc. These verbs do not convey any additional information while generating the topics. We improve the trigram phrases that contain the progressive verb at the first or last position of the phrases by removing the progressive verbs from the trigram phrases (rem operation). We present rules as follows:

$$w1\langle VBG \rangle w2\langle T \rangle w3\langle T \rangle \xrightarrow{\text{rem}(w1)} w2\langle T \rangle w3\langle T \rangle$$

$$w1\langle T \rangle w2\langle T \rangle w3\langle VBG \rangle \xrightarrow{\text{rem}(w3)} w1\langle T \rangle w2\langle T \rangle$$

By using the above rules, phrases like ‘neural network modeling’, ‘solving linear system’ are transformed into ‘neural network’, ‘linear system’ respectively. These shorter phrases without progressive verb form better topics that can convey a more general concept.

### 7.2.5 Topic Grouping

To group similar topics, we find similarities among the topics by exploiting their co-occurrences in the papers. We use Google’s Word2vec model [113] to find the similarity among the topics. Word2vec model creates the word-embeddings by generating vector space from the text corpus where each word in the corpus is assigned to a vector in the space. As abstracts have more contextual information, we use abstracts [164] of the papers to find similar topics. We first train the Word2vec model using abstracts of the papers and then use  $k$ -medoid algorithm [165] to group the topics generated by the proposed method based on their Word2vec similarities. We use  $k$ -medoid algorithm instead of widely used  $k$ -means algorithm because of its robustness to outliers as compared to  $k$ -means.

## 7.3 Experimental Evaluations

In this section, we present our evaluation. Section 7.3.1 describes the experimental setup. In Section 7.3.2, we compare the performance of all the algorithms using precision and similarity measures. In Section 7.3.3, we show through empirical results that the topics generated by our proposed method are superior to existing popular topic modeling algorithms.

### 7.3.1 Experimental Setup

We use publicly available DBLP bibliographic dataset<sup>4</sup> for our experiments. The dataset contains information from more than 5000 conferences, 1500 journals, and indexes over 3.3 million publications of Computer Science. We also extract data from WikiCFP to categorize the publications based on research areas. WikiCFP is a platform that lists conferences, scholarly events, meetings and allows for advertising

---

<sup>4</sup><http://dblp.uni-trier.de/>

information about workshops, conferences, seminars, meetings, etc. It includes 276 research categories, most of which are from Computer Science. We extract the data from WikiCFP to know the conferences that belong to different areas of Computer Science, and then select 27 most popular areas, such as databases, machine learning, etc., for our analysis.

As the dataset contains many noisy and unimportant words, we do pre-processing to remove these words. It reduces the time requirement of the phrase discovery process and the search space. We remove stop-words, such as ‘a’, ‘an’, ‘the’, etc., as these words do not contain significant information for our analysis. Stemming and lemmatization are two commonly used text pre-processing techniques. These techniques reduce inflected or derived words to their root forms. However, we found that these techniques create many distorted topics. For example, if we apply these pre-processing, we get topics as *distribut databas system* instead of *distributed database system*. We observed that paper titles, unlike paper abstracts, often contain good topics and they do not have inflected words. In this chapter, we do very mild stemming to remove redundant topics. For example, we found redundant topics such as Database and Databases; Neural Network and Neural Networks. For such redundant topics, we chose the topic that was more frequent in the dataset.

### 7.3.2 Performance Evaluation

In this section, we compare the performance of different topic modeling algorithms. There are different metrics to evaluate topic models such as perplexity, precision, similarity, etc. Chang et. al [139] showed that the perplexity metric is not well suited for topic model evaluation as it is not well correlated with human judgment. The topic models that achieve better perplexity have less human-interpretable topics. Therefore, we use precision and similarity metrics for evaluations. Later in Section 7.3.3, we will show the superiority of our generated topics using empirical results.

## Precision Analysis

In topic modeling, precision is defined as the fraction of generated topics that are relevant to the topics of interest, which could either be of a conference or of a research area. Since there is no labeled data for this evaluation, we asked 10 PhD researchers from our CSE department to manually label the topical words and phrases (topics) generated by the topic finding algorithms. We chose these researchers from different areas of Computer Science. We chose two researchers, with expertise in the same area, to label the topics generated from research area or conference. Topics of each conference and area are labeled by researchers independently without influencing each other. We consider a topic as a relevant topic if both the researchers labeled it as a relevant topic. We found that there was 94% agreement among the researchers while labeling the topics. Topics for which researchers did not agree on were discussed until a consensus was reached. Due to a dearth of experts from different domains, it is difficult to find researchers from several areas to judge the results. We therefore perform our experiments on 20 conferences and 10 areas. We compute the mean average precision (MAP) [140] for each algorithm by averaging the precision values of the conferences and areas. In the following table, ‘w/’ indicates ‘with’.

<b>Techniques/ MAP(%)</b>	<b>Baseline</b>	<b>w/ Pruning</b>	<b>w/ Refining</b>	<b>w/ Pruning + Refining</b>
LDA	51.4	56.7	51.4	56.7
TNG	68.3	72.3	70.1	75.2
Proposed Method	62.8	79.4	68.9	89.1

Table 7.1: MAP of different topic modeling algorithms

As we discussed in Section 7.2, our proposed method consists of several NLP processing. We compare the performance of all the methods with and without NLP processing, namely the pruning and the refining step. We compute the precision of topic finding algorithms with pruning, with refining and with both pruning and

refining. From the first row of Table 7.1, we observe that LDA has the lowest MAP value (51.4%). The reason for this poor result is that LDA generates single words and many of them do not convey any information as these words are part of topic phrases. LDA with refining does not improve the precision as refining does not apply to LDA single terms, while LDA with pruning performs better (56.7%).

TNG performs better than LDA, it gives 75.2% MAP with both pruning and refining but it generates many irrelevant words and phrases that lower the performance of TNG. The performance of proposed method without any processing is less compared to TNG. The reasons for this is that many of the topics are disordered, redundant and incomplete. However, when topic pruning and topic refining are individually added to the proposed method, there are notable improvements, but the largest improvement comes when topics generated by association mining are processed by both pruning and refining, yielding 89.1% mean average precision. The reason for this is that after processing the topics, the proposed method generates relatively more meaningful as well as relevant topics compared to other methods. Further, applying pruning on the list of generated topics decreases recall of the proposed method. However, we set a fitting value of support such that the proposed method generates adequate topics (as shown in Table 7.3).

## Similarity Analysis

We next use similarity measures to evaluate the performance of different topic finding algorithms. We want to compare how the topics generated by the algorithms compare with manually listed topics, similar to the one shown in Figure 7.1. However, the challenge is that there is no existing gold-standard list of topics to compare against the generated topics. We created our baseline topics for conferences and research areas by collecting topics from Wikipedia and WikiCFP.

Wikipedia is a publicly available encyclopedia which provides domain-specific in-

formation. We extract topics of conferences and areas using Wikipedia API<sup>5</sup>. However, Wikipedia offers much more information than required for analysis of topics. In order to avoid adding noise, we only consider anchor texts present in the infobox table and first two paragraphs from the searched web page as this is the most related information. We also extract the list of the topics by crawling WikiCFP with the name of conference and area which are provided by the conferences to call for papers<sup>6</sup>. Conferences contain the topics of their interests but areas contain the conferences belong to the area. To get the topics of the area, we take the topics of top-20 conferences of that area. We then compute the cosine similarity between the collected topics and topics generated by the algorithms. We present the similarity score of a few popular areas in Table 7.2.

<b>Algorithms /Results</b>	<b>LDA</b>	<b>TNG</b>	<b>Proposed Method</b>
Machine Learning	0.14	0.26	0.51
Computer Networks	0.16	0.25	0.45
Compilers	0.13	0.16	0.38
Databases	0.10	0.21	0.36

Table 7.2: Similarity of topics with manually listed topics

As can be seen in Table 7.2, LDA has the lowest similarity for all the areas. Although TNG performs better than LDA, its similarity is also low. The reason for this low similarity is that most of the topics generated by these algorithms do not match with manually written topics as collected from WikiCFP and Wikipedia. However, the proposed algorithm with pruning and refining performs the best, with almost twice higher similarity compared to TNG. Although the absolute values of similarity appear small, they account for a large number of topics, and many of them are related but do not match exactly. The cosine similarity is incapable of matching the topics if they are related but have different terms. Further, we observe that similarity varies across

<sup>5</sup><https://pypi.python.org/pypi/wikipedia>

<sup>6</sup>Earlier in Section 7.2.1, we used WikiCFP to get the name of conferences that belong to a research area unlike the current procedure to collect the topics.



the areas for all the methods as it is highly dependent on the quality of the baseline topics generated through Wikipedia and WikiCFP. We therefore collect the topics of 20 conferences and 10 research areas of Computer Science and compute the average similarity score. We obtain the average similarity scores of the topics generated by LDA, TNG and proposed algorithm as 0.127, 0.228, 0.435 respectively. These results are inline with the results shown in Table 7.1 reporting that the proposed algorithm performs better than LDA and TNG.

### 7.3.3 Empirical Analysis

In this section, we compare different topic finding algorithms using empirical evaluation. We then show a sample application of using our topic finding algorithm over bibliographic data.

#### Topic Comparison

In Table 7.3, we show the topics generated by LDA, TNG and our proposed for the *Machine Learning* research area.

As can be seen in Table 7.3, LDA generates topics that are generic words such as ‘model’, ‘data’, ‘algorithm’, ‘approach’, ‘control’, etc. Many of them are not actual topics and some of them are part of topic phrases, such as ‘learning’, ‘data’, and ‘neural’ are the parts of ‘reinforcement learning’, ‘big data’, and ‘neural network’ respectively. Many topic words such as ‘based’, ‘method’, ‘control’ do not convey any meaningful information. In general, phrases convey more useful information compared to single words. Most of the manually created topics are also phrases. TNG generates phrases but many top phrases of TNG are noisy and less interpretable such as ‘based divisive’, ‘based representability’, ‘case study’, ‘empirical study’, etc. Even after we remove the terms like ‘based’, ‘study’, ‘approach’ by pre-processing, TNG generates many not so useful topics such as ‘visualize high’, ‘large scale’, ‘logical

<b>LDA</b>	<b>TNG</b>	<b>Proposed Method</b>
learning	neural approach	neural network
based	visualize high	reinforcement learning
data	automatic translation	face recognition
neural	logical foundation	swarm optimization
network	based representability	matrix factorization
algorithm	based divisive	semi-supervised learning
model	dimensional shape	big data
analysis	investigating temporal	recommender system
system	channel state	association rule
approach	neural network	monte carlo
classification	genetic programming	social media
detection	active learning	k-means clustering
recognition	empirical study	multi-label classification
image	large scale	logistic regression
clustering	support vector	sentiment analysis
mining	data mining	support vector machine
control	big data	artificial neural network
genetic	online learning	hidden markov model
method	vectorial data	principal component analysis
optimization	feature selection	deep neural network
information	case study	natural language processing

Table 7.3: Key-topics of Machine Learning area

foundation’, etc. Without the NLP pruning and refining, our proposed approach also generates many tedious and less interpretable topics such as ‘using support vector’, ‘based association rule’, ‘processing natural language’, ‘neural networks training’, etc. However, after performing pruning and refining text processing, our proposed method generates topics which are quite similar to manually written topics.

### Topic Evolution in Databases

In Table 7.4, we show topics of interest evolution in Databases for three decades. These are the topics obtained from the publications in Databases during the years 1985 – 1994, 1995 – 2004 and 2005 – 2014. Any Database researcher can confirm that the topic evolution shown in Table 7.4 closely matches the reality. We have shown top-20 topics from each period. However, if we show more number of topics

1985-1995	1995-2005	2005-2015
concurrency control	world wide web	big data
query optimization	data reduction	anomaly detection
relational database system	decision support system	matrix factorization
database programming language	large database	social network
relational algebra	mining frequent pattern	intrusion detection
relational database management	high dimensional data	dimensionality reduction
distributed concurrency control	conjunctive queries	natural language
query processing database	information system development	keyword search
optimistic concurrency control	decision support system	unstructured data
knowledge discovery	expert system	knowledge base
heterogeneous database	data warehouse	data cubes
multiprocessor database	query refinement	time series data
reverse engineering	search engines	particle swarm optimization
query language	knowledge management system	deep web
database system programming	distributed query processing	business intelligence
deadlock detection	multimedia database system	learning rank
spatial database	data streams	semantic web services
knowledge discovery database	data integration system	user interaction
crash recovery	privacy preserving	cloud platform
transaction execution	digital library	relevance query

Table 7.4: Topic evolution in Databases over the years

then even less popular topics of interest, such as Skyline Queries, Database Usability, Provenance, etc., will also show up in the result. Getting this kind of topic evolution information can be very useful to researchers who want to explore other research areas. For example, if someone is in Databases and wants to explore topics in NLP, one will find it very useful to know how the research in NLP has evolved over the years. One can know the ongoing or the old topics in a research area or a conference.

As can be seen from Table 7.4, in the interval 1985 – 1995 the database community has focused mainly on different data models, database design and query optimization. During 1995 – 2005 due to rapid advancement in world wide web and search engines, the community started to focus on managing data generated from the web, such as integrating different heterogeneous data sources, streaming data in the form of click streams, etc. Frequent pattern mining was another major research topic during this period. During 2005-2015, the focus has shifted to social network, natural language processing, keyword search, big data, etc.

## Grouping of the topics

To provide a comprehensive summary of topics of interest to a conference or a research area, we create groups of similar topics. These groups provide the overall theme of a conference or a research area. Table 7.5 presents the groups that are obtained from *Machine Learning* research area.

Groups	Topics
Group1	neural network, deep neural network, fuzzy neural network, genetic algorithm, handwritten recognition, artificial neural network
Group2	natural language, sentiment analysis, knowledge bases, data streams, social media, big data
Group3	support vector classification, support vector regression, active learning, face recognition, feature selection
Group4	hidden markov model, conditional random fields, decision making, ant colony, monte carlo, reinforcement learning
Group5	matrix factorization, recommender system, collaborative filtering, dimensionality reduction, principal component analysis

Table 7.5: Topic Groups of Machine Learning research

As can be seen in Table 7.5, topics, which are related, are placed under the same group. Group 1 consists of topics related to neural network such as deep neural network, artificial neural network, fuzzy neural network, handwritten recognition, etc. Similarly, Group 2 consists of topics related to natural language analysis such as sentiment analysis, knowledge bases, data streams, social media, and so on. Each group contains a set of semantically related topics.

## 7.4 Conclusion

In this chapter, we proposed a novel information summarization method based on association mining and natural language processing to generate topics of interest. We used our method on bibliographic data to generate topics of interest for conferences and research areas. We showed that the proposed method can generate topics, which

are very similar to manually created topics of interest. We implemented various NLP based pruning and refining techniques to get near-perfect topics. The experimental results showed that the proposed method generated topics with higher precision compared to probabilistic topic modeling algorithms. Finally, we showed in our evaluation results that the topics generated by our proposed method are more meaningful and human interpretable than the existing state-of-the-art methods.

# Chapter 8

## Conclusion and Future Work

In this chapter, we first give a summary of the thesis and later wrap it up by pointing out some possible future directions.

### 8.1 Summary of the Thesis

In this thesis, we studied the problem of information diffusion and information summarization in social networks. We analyzed three important factors, namely network connectivity, posting time and post content, which are crucial factors of information diffusion in social networks. To summarize huge volumes of social network text data, we propose a novel topic generation algorithm based on NLP and frequent pattern mining algorithm. Our generated topics are more similar to manually created topics than existing topic modeling algorithms. A brief summary of all the proposed solutions is presented in the following subsections.

#### 8.1.1 Information Diffusion using WoM Marketing

Network connectivity is one of the most important determinants of information diffusion. We proposed a method to do widespread WoM marketing in OSGs using

network connectivity. An authoritative user of the network can disseminate information to a more wider audience compared to an ordinary user. To find authoritative users, it is essential to build a social interaction graph using social interactions such as posts, likes, comments, likes-on-comment, shares, etc. We presented a novel algorithm to build a social interaction graph. As influence of a user varies across the topics, we created topic-sensitive social interaction graph. Using link analysis from social network analysis, we found the topic-sensitive authoritative users of the network. These prominent authoritative users are influential in the topic. Commercial organizations can use these users to do WoM marketing in the network. We next presented a concept of reinforced WoM marketing, where multiple authorities can together promote a product to increase the effectiveness of marketing. Finally, we showed the best time of the year to start marketing in OSGs to further improve the performance of the marketing.

### **8.1.2 Information Diffusion using the Best Time to Post**

In Chapter 4, we looked at the problem of finding the best posting time(s) to get high information diffusion. We did our analysis over Facebook pages and revealed that most of the reactions are received within a few hours of posting. To maximize the diffusion of a post by increasing audience reactions, we introduced six posting schedules that can be used for individual pages or group of pages with similar audience reaction profile. Our best posting schedule can lead to seven times more number of audience reactions compared to the average number of audience reactions that users would get without following any optimized posting schedule. We presented some interesting audience reaction patterns that we obtained through daily, weekly and monthly audience reaction analysis. Finally, we showed types of content that receive more audience reactions. Pages can achieve higher audience engagement by creating contents of the type that receives more audience reactions.

### **8.1.3 Information Diffusion by Posting High Arousal Content**

In Chapter 5, we predicted news posts that have high potential to generate high-arousal. High arousal posts would attract a large number of users to give their opinions in the form of comments. We predicted the arousal of news posts prior to their release, which brings the possibility of appropriate decision making to modify the post content or its ranking in audience newsfeed. We generated multiple features from the content of news posts and showed that our best set of features with feature selection can predict the arousal with high accuracy. We also showed the topics of high arousal that can be included in news posts to achieve high arousal.

### **8.1.4 Sentiment Dynamics in Social Media Channels**

Social media is presently one of the most important means of news communication. We analyzed how news channels use sentiment to garner users' attention in social media. We compared the sentiment of social media news posts of television, print media and radio, to show the variations in the ways these channels report the news. We also analyzed users' reactions and users' opinions on news posts having different sentiments. We performed our experiments on a dataset extracted from five popular Facebook pages of three different types of news channels. Experimental results showed that the sentiment of news posted by different types of social media channels is dependent on the medium through which these channels traditionally disseminated news and news with positive or negative sentiment receive lots of users' attention. We also revealed that the sentiment of user opinion has a strong correlation with the type of information source and the sentiment of the news post.

### **8.1.5 Information Summarization by Generating Key-topics**

In Chapter 7, we proposed a method for information summarization by generating topics of interest from large social network research publications using association



mining and NLP. We implemented various NLP based pruning and refining techniques to get well-formed, meaningful and interpretable topics. We showed that the topics that are generated by the proposed method are much more precise and similar to manually written topics of interest compared to existing topic finding algorithms. We also presented a comprehensive summary of topics by grouping these topics using clustering based on semantic similarities.

## 8.2 Future Work

In this section, we give a few of the many possible future directions that have been opened up by this thesis.

1. To find topic-sensitive WoM marketers, we create a topic-sensitive social interaction graph from users' interactions in OSGs. Although the topic-sensitive graph captures the topic-sensitivity of users of the network, it does not create embeddings of nodes preserving the local neighborhood in the space. In our future work, we would create node-embeddings from the network interactions such that all the topically related similar users would have similar embeddings in the vector space [166, 167]. The graph can be used for better categorization of networks and efficient retrieval such as finding similar users for a given user.
2. To find the best time to post, we focus on the pages of the same location. In our future work, we can look at how to aggregate the pages of different locations and perform a similar analysis. One can also look into the effect of post topic on audience reactions. Further, individuals and commercial organizations can use temporal patterns from our analysis to predict future trends using time series analysis [168, 169].
3. Predicting arousal of social media contents is a very compelling problem. There is a high scope to extend the problem as many studies in this direction have

not been conducted so far. One can use different features and models such as recurrent neural network, long short-term memory [170] to perform arousal prediction and can compare with our existing method.

4. While studying sentiment dynamics in social media, we have done our analysis on Facebook pages of traditional news channels such as TV, print-media, and radio-based channels. However, we did not study the sentiment dynamics of actual news articles published in the websites of these channels. In our future work, we would like to include actual news stories from different types of channels and would compare them with social media news posts. We would also compare news posted by these channels through their microblogging handles.
5. We summarize social network text corpus by generating topics of interest. One of the fundamental problems in social media text summarization is that social media texts are usually very short in length and context information is not easily available that could describe these texts in detail. Therefore, generating tags or topics for these short texts is a challenging problem. One of the possible extension is to aggregate similar texts in an efficient way before applying a topic finding algorithm to generate topics. Another possible future work can be carried out by adding more post-processing and NLP based refining steps to get more well-formed topics similar to manually created topics.

# Bibliography

- [1] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.
- [2] C.-T. Li, Y.-J. Lin, and M.-Y. Yeh, “Forecasting participants of information diffusion on social networks with its applications,” *Information Sciences*, vol. 422, pp. 432–446, 2018.
- [3] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter,” in *WWW*. ACM, 2011, pp. 695–704.
- [4] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, “Who says what to whom on twitter,” in *WWW*. ACM, 2011, pp. 705–714.
- [5] M. Trusov, A. V. Bodapati, and R. E. Bucklin, “Determining influential users in internet social networks,” *Journal of Marketing Research*, vol. 47, no. 4, p. 643–658, 2010.
- [6] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” in *WSDM*. ACM, 2010, pp. 261–270.
- [7] N. Kumar, Y. Chandarana, K. Anand, and M. Singh, “Using social media for word-of-mouth marketing,” in *International Conference on Big Data Analytics and Knowledge Discovery*. Springer, 2017, pp. 391–406.

- [8] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, “Trends in social media: Persistence and decay,” in *ICWSM*, 2011, pp. 434–437.
- [9] N. Kumar, G. Ande, J. S. Kumar, and M. Singh, “Toward maximizing the visibility of content in social media brand pages: a temporal analysis,” *Social Network Analysis and Mining*, vol. 8, no. 1, p. 11, 2018.
- [10] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, “Bad news travel fast: A content-based analysis of interestingness on twitter,” in *Proceedings of the 3rd International Web Science Conference*. ACM, 2011, p. 8.
- [11] N. Kumar, A. Yadandla, K. Suryamukhi, N. Ranabothu, S. Boya, and M. Singh, “Arousal prediction of news articles in social media,” in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2017, pp. 308–319.
- [12] N. Kumar, R. Nagalla, T. Marwah, and M. Singh, “Sentiment dynamics in social media news channels,” *Online Social Networks and Media*, vol. 8, pp. 42–54, 2018.
- [13] J. Prüfer and P. Prüfer, “Data science for entrepreneurship research: Studying demand dynamics for entrepreneurial skills in the netherlands,” *CentER Discussion Paper, 2019-005*, p. 36, 2019.
- [14] N. Kumar, R. Utkoor, B. K. Appareddy, and M. Singh, “Generating topics of interests for research communities,” in *International Conference on Advanced Data Mining and Applications*. Springer, 2017, pp. 488–501.
- [15] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” in *WWW*. ACM, 2012, pp. 519–528.

- [16] K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Super mediator—a new centrality measure of node importance for information diffusion over social network,” *Information Sciences*, vol. 329, pp. 985–1000, 2016.
- [17] Z. Wang, Y. Yang, J. Pei, L. Chu, and E. Chen, “Activity maximization by effective information diffusion in social networks,” *TKDE*, vol. 29, no. 11, pp. 2374–2387, 2017.
- [18] K. Shu, H. R. Bernard, and H. Liu, “Studying fake news via network analysis: detection and mitigation,” in *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer, 2019, pp. 43–65.
- [19] V. Arnaboldi, M. Conti, A. Passarella, and R. I. Dunbar, “Online social networks and information diffusion: The role of ego networks,” *Online Social Networks and Media*, vol. 1, pp. 44–55, 2017.
- [20] J. Yang and J. Leskovec, “Modeling information diffusion in implicit networks,” in *ICDM*. IEEE, 2010, pp. 599–608.
- [21] J. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,” *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.
- [22] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *SIGKDD*. ACM, 2003, pp. 137–146.
- [23] M. Kimura, K. Saito, and H. Motoda, “Blocking links to minimize contamination spread in a social network,” *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 2, p. 9, 2009.

- [24] M. Granovetter, “Threshold models of collective behavior,” *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.
- [25] D. J. Watts, “A simple model of global cascades on random networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5766–5771, 2002.
- [26] D. J. Watts and P. S. Dodds, “Influentials, networks, and public opinion formation,” *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.
- [27] D. Centola, “The spread of behavior in an online social network experiment,” *science*, vol. 329, no. 5996, pp. 1194–1197, 2010.
- [28] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?” in *WWW*. ACM, 2014, pp. 925–936.
- [29] M. Farajtabar, X. Ye, S. Harati, L. Song, and H. Zha, “Multistage campaigning in social networks,” in *NIPS*, 2016, pp. 4718–4726.
- [30] A. Guille, “Information diffusion in online social networks,” in *Proceedings of the 2013 SIGMOD/PODS Ph. D. symposium*. ACM, 2013, pp. 31–36.
- [31] K. Lerman and R. Ghosh, “Information contagion: An empirical study of the spread of news on digg and twitter social networks.” *ICWSM*, vol. 10, pp. 90–97, 2010.
- [32] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, “Patterns of cascading behavior in large blog graphs,” in *SDM*. SIAM, 2007, pp. 551–556.
- [33] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, “Prominent features of rumor propagation in online social media,” in *ICDM*. IEEE, 2013, pp. 1103–1108.

- [34] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in *SIGKDD*. ACM, 2009, pp. 497–506.
- [35] S. Pei and H. A. Makse, “Spreading dynamics in complex networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, no. 12, p. P12002, 2013.
- [36] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.
- [37] M. Tsytsarau, T. Palpanas, and M. Castellanos, “Dynamics of news events and social media reaction,” in *SIGKDD*. ACM, 2014, pp. 901–910.
- [38] D. Vogiatzis, “Influential users in social networks,” in *Semantic Hyper/Multimedia Adaptation*. Springer, 2013, pp. 271–295.
- [39] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *SIGKDD*. ACM, 2009, pp. 199–208.
- [40] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” in *ICWSM*, 2010, pp. 10–17.
- [41] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *WWW*. ACM, 2010, pp. 591–600.
- [42] J. Zhang, M. S. Ackerman, and L. Adamic, “Expertise networks in online communities: structure and algorithms,” in *WWW*. ACM, 2007, pp. 221–230.
- [43] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: bringing order to the web,” 1999.
- [44] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, “The web as a graph: measurements, models, and methods,” in *International Computing and Combinatorics Conference*. Springer, 1999.

- [45] B. Ruhnau, “Eigenvector-centrality—a node-centrality?” *Social networks*, vol. 22, no. 4, 2000.
- [46] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, 1977.
- [47] Freeman and L. C., “Centrality in social networks conceptual clarification,” *Social networks*, 1978.
- [48] Forbes. (2013) What are they saying about your brand? <http://www.forbes.com/sites/pauljankowski/2013/02/27/quick-what-are-they-saying-about-your-brand/ee5ff7374a8d>.
- [49] J. Li, W. Peng, T. Li, T. Sun, Q. Li, and J. Xu, “Social network user influence sense-making and dynamics prediction,” *Expert Systems with Applications*, vol. 41, no. 11, pp. 5115–5124, 2014.
- [50] A. Das, S. Gollapudi, and K. Munagala, “Modeling opinion dynamics in social networks,” in *WSDM*. ACM, 2014, pp. 403–412.
- [51] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, “Dynamical classes of collective attention in twitter,” in *WWW*. ACM, 2012, pp. 251–260.
- [52] S. Wu, A. Das Sarma, A. Fabrikant, S. Lattanzi, and A. Tomkins, “Arrival and departure dynamics in social networks,” in *WSDM*. ACM, 2013, pp. 233–242.
- [53] L. Yu, P. Cui, C. Song, T. Zhang, and S. Yang, “A temporally heterogeneous survival framework with application to social behavior dynamics,” in *SIGKDD*. ACM, 2017, pp. 1295–1304.
- [54] M. R. Karimi, E. Tavakoli, M. Farajtabar, L. Song, and M. Gomez Rodriguez, “Smart broadcasting: Do you want to be seen?” in *SIGKDD*. ACM, 2016, pp. 1635–1644.



- [55] N. Spasojevic, Z. Li, A. Rao, and P. Bhattacharyya, “When-to-post on social networks,” in *SIGKDD*. ACM, 2015, pp. 2127–2136.
- [56] A. Zarezade, U. Upadhyay, H. R. Rabiee, and M. Gomez-Rodriguez, “Redqueen: An online algorithm for smart broadcasting in social networks,” in *WSDM*. ACM, 2017, pp. 51–60.
- [57] R. Biswas, D. Riffe, and D. Zillmann, “Mood influence on the appeal of bad news,” *Journalism & Mass Communication Quarterly*, vol. 71, no. 3, pp. 689–696, 1994.
- [58] C. Esiyok, B. Kille, B.-J. Jain, F. Hopfgartner, and S. Albayrak, “Users’ reading habits in online news portals,” in *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, 2014, pp. 263–266.
- [59] J. C. S. dos Rieis, F. B. de Souza, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An, “Breaking the news: First impressions matter on online news,” in *ICWSM*, 2015, pp. 357–366.
- [60] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, “To better stand on the shoulder of giants,” in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. ACM, 2012, pp. 51–60.
- [61] D. G. Taylor, J. E. Lewin, and D. Strutton, “Friends, fans, and followers: do ads work on social networks?” *Journal of advertising research*, vol. 51, no. 1, pp. 258–275, 2011.
- [62] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, “Predicting the popularity of online articles based on user comments,” in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. ACM, 2011, p. 67.

- [63] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, “A peek into the future: Predicting the evolution of popularity in user generated content,” in *WSDM*. ACM, 2013, pp. 607–616.
- [64] A. Tatar, P. Antoniadis, M. D. De Amorim, and S. Fdida, “From popularity prediction to ranking online news,” *Social Network Analysis and Mining*, vol. 4, no. 1, p. 174, 2014.
- [65] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, “Predicting information spreading in twitter,” in *Workshop on computational social science and the wisdom of crowds, NIPS*. Citeseer, 2010, pp. 17 599–601.
- [66] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in *IEEE Second International Conference on Social Computing*. IEEE, 2010, pp. 177–184.
- [67] S. Petrovic, M. Osborne, and V. Lavrenko, “Rt to win! predicting message propagation in twitter.” in *ICWSM*, vol. 11, 2011, pp. 586–589.
- [68] L. Weng, F. Menczer, and Y.-Y. Ahn, “Predicting successful memes using network and community structure.” in *ICWSM*, vol. 8, 2014, pp. 535–544.
- [69] J. G. Lee, S. Moon, and K. Salamatian, “An approach to model and predict the popularity of online contents with explanatory factors,” in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 623–630.
- [70] S. Wu, C. Tan, J. M. Kleinberg, and M. W. Macy, “Does bad news go away faster?” in *ICWSM*, 2011, pp. 646–649.
- [71] R. Bandari, S. Asur, and B. A. Huberman, “The pulse of news in social media: Forecasting popularity.” in *ICWSM*, 2012, pp. 26–33.

- [72] J. Reis, P. Gonçalves, P. Vaz de Melo, R. Prates, and F. Benevenuto, “Magnet news: You choose the polarity of what you read,” in *ICWSM*, 2014, pp. 652–653.
- [73] J. C. S. dos Reis, F. B. de Souza, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An, “Breaking the news: First impressions matter on online news,” in *ICWSM*, 2015, pp. 357–366.
- [74] A. Zubiaga, “Newspaper editors vs the crowd: on the appropriateness of front page news selection,” in *WWW. ACM*, 2013, pp. 879–880.
- [75] T. Demeester, D. Nguyen, D. Trieschnigg, C. Develder, and D. Hiemstra, “What snippets say about pages in federated web search,” in *Asia Information Retrieval Symposium*. Springer, 2012, pp. 250–261.
- [76] D. Wightman, Z. Ye, J. Brandt, and R. Vertegaal, “Snipmatch: using source code context to enhance snippet retrieval and parameterization,” in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 219–228.
- [77] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Enhancing topic modeling for short texts with auxiliary word embeddings,” *ACM Transactions on Information Systems*, vol. 36, no. 2, p. 11, 2017.
- [78] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, “Topic modeling of short texts: A pseudo-document view,” in *SIGKDD*. ACM, 2016, pp. 2105–2114.
- [79] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, “Context preserving dynamic word cloud visualization,” in *Visualization Symposium (Pacific Vis), 2010 IEEE Pacific*. IEEE, 2010, pp. 121–128.

- [80] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [81] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in *ICDM*. IEEE, 2007, pp. 697–702.
- [82] W. Li and A. McCallum, “Pachinko allocation: Dag-structured mixture models of topic correlations,” in *ICML*. ACM, 2006, pp. 577–584.
- [83] Q. Mei, C. Liu, H. Su, and C. Zhai, “A probabilistic approach to spatiotemporal theme pattern mining on weblogs,” in *WWW*. ACM, 2006, pp. 533–542.
- [84] Q. Mei and C. Zhai, “A mixture model for contextual text mining,” in *SIGKDD*. ACM, 2006, pp. 649–655.
- [85] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, “Probabilistic author-topic models for information discovery,” in *SIGKDD*. ACM, 2004, pp. 306–315.
- [86] T. Hofmann, “Probabilistic latent semantic indexing,” in *SIGIR*. ACM, 1999, pp. 50–57.
- [87] H. M. Wallach, “Topic modeling: beyond bag-of-words,” in *ICML*. ACM, 2006, pp. 977–984.
- [88] R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic, “A phrase-discovering topic model using hierarchical pitman-yor processes,” in *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*. ACL, 2012, pp. 214–222.
- [89] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, “Scalable topical phrase mining from text corpora,” *VLDB*, vol. 8, no. 3, pp. 305–316, 2014.

- [90] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, “Graph structure in the web,” *Computer networks*, vol. 33, no. 1, pp. 309–320, 2000.
- [91] J. R. Firth, “A synopsis of linguistic theory, 1930-1955,” *Studies in linguistic analysis*, 1957.
- [92] P. E. Latham and Y. Roudi, “Mutual information,” *Scholarpedia*, vol. 4, no. 1, p. 1658, 2009.
- [93] T. Bucher, “Want to be on the top? algorithmic power and the threat of invisibility on facebook,” *new media & society*, vol. 14, no. 7, pp. 1164–1180, 2012.
- [94] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.
- [95] Facebook. (2019) Facebook graph api. <https://developers.facebook.com/docs/graph-api>.
- [96] J. Perkins, *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd, 2014.
- [97] I. Lawrence and K. Lin, “A concordance correlation coefficient to evaluate reproducibility,” *Biometrics*, 1989.
- [98] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.
- [99] M. Levene, *Web dynamics: Adapting to change in content, size, topology and use*. Springer Science & Business Media, 2004.

- [100] L. Backstrom. (2013) News feed fyi: A window into news feed. <https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed>.
- [101] J. Weaver and P. Tarjan, “Facebook linked data via the graph api,” *Semantic Web*, vol. 4, no. 3, pp. 245–250, 2013.
- [102] M. Mazloom, R. Rietveld, S. Rudinac, M. Worrying, and W. van Dolen, “Multi-modal popularity prediction of brand-related social media posts,” in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 197–201.
- [103] C. Kim and S.-U. Yang, “Like, comment, and share on facebook: How each behavior differs from the other,” *Public Relations Review*, vol. 43, no. 2, pp. 441–449, 2017.
- [104] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, “K nearest neighbours with mutual information for simultaneous classification and missing data imputation,” *Neurocomputing*, vol. 72, no. 7, pp. 1483–1493, 2009.
- [105] U. M. Fayyad and K. B. Irani, “On the handling of continuous-valued attributes in decision tree generation,” *Machine learning*, vol. 8, no. 1, pp. 87–102, 1992.
- [106] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.
- [107] A. McCallum, K. Nigam *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

- [108] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in k-means clustering,” *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.
- [109] N. Diakopoulos and M. Naaman, “Topicality, time, and sentiment in online news comments,” in *CHI’11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011, pp. 1405–1410.
- [110] F. Figueiredo, H. Pinto, F. Belém, J. Almeida, M. Gonçalves, D. Fernandes, and E. Moura, “Assessing the quality of textual features in social media,” *Information Processing & Management*, vol. 49, no. 1, pp. 222–247, 2013.
- [111] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit.” in *ACL (System Demonstrations)*, 2014, pp. 55–60.
- [112] D. M. Christopher, R. Prabhakar, and S. Hinrich, “Introduction to information retrieval,” *An Introduction To Information Retrieval*, vol. 151, p. 177, 2008.
- [113] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [114] L. Lusa *et al.*, “Smote for high-dimensional class-imbalanced data,” *BMC bioinformatics*, vol. 14, no. 1, p. 106, 2013.
- [115] T. G. Dietterich, “Ensemble learning,” *The handbook of brain theory and neural networks*, vol. 2, pp. 110–125, 2002.
- [116] L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

- [117] C. Manning, “Information extraction and named entity recognition,” Stanford University Lecture, 2012.
- [118] S. Bengio, J. Mariéthoz, and M. Keller, “The expected performance curve,” in *International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, no. EPFL-CONF-83266, 2005.
- [119] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, “The effects of subtle misinformation in news headlines.” *Journal of experimental psychology: applied*, vol. 20, no. 4, p. 323, 2014.
- [120] S. Soroka and S. McAdams, “News, politics, and negativity,” *Political Communication*, vol. 32, no. 1, pp. 1–22, 2015.
- [121] J. Gottfried and E. Shearer, “News use across social media platforms 2016,” *Pew Research Center*, vol. 26, 2016.
- [122] P. Rozin and E. B. Royzman, “Negativity bias, negativity dominance, and contagion,” *Personality and social psychology review*, vol. 5, no. 4, pp. 296–320, 2001.
- [123] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.
- [124] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: a survey,” *International Journal*, vol. 2, no. 6, pp. 282–292, 2012.
- [125] M. Bautin, L. Vijayarenu, and S. Skiena, “International sentiment analysis for news and blogs,” in *ICWSM*, 2008, pp. 19–26.
- [126] N. Godbole, M. Srinivasaiah, and S. Skiena, “Large-scale sentiment analysis for news and blogs,” *ICWSM*, vol. 7, no. 21, pp. 219–222, 2007.



- [127] T. E. Patterson, *Out of Order: An incisive and boldly original critique of the news media's domination of America's political process*. Vintage, 2011.
- [128] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.
- [129] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing & Management*, vol. 50, no. 1, pp. 104–112, 2014.
- [130] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, 2014, pp. 216–225.
- [131] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, vol. 2017. NIH Public Access, 2017, p. 1217.
- [132] T. Davidson, D. Warmusley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *ICWSM*, 2017, pp. 512–515.
- [133] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *LREC*, vol. 10, no. 2010, 2010, pp. 2200–2204.
- [134] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis." in *FLAIRS conference*, 2012, pp. 202–207.
- [135] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, 2001.

- [136] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004.
- [137] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, “A heuristic approach to determine an appropriate number of topics in topic modeling,” *BMC bioinformatics*, vol. 16, no. 13, p. S8, 2015.
- [138] R. E. Madsen, D. Kauchak, and C. Elkan, “Modeling word burstiness using the Dirichlet distribution,” in *ICML*. ACM, 2005, pp. 545–552.
- [139] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, “Reading tea leaves: How humans interpret topic models.” in *NIPS*, vol. 31, 2009, pp. 1–9.
- [140] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, “Diversifying search results,” in *WSDM*. ACM, 2009, pp. 5–14.
- [141] C. Sievert and K. E. Shirley, “Ldavis: A method for visualizing and interpreting topics,” in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.
- [142] W. L. Bennett, *News: The politics of illusion*. University of Chicago Press, 2016.
- [143] L. K. Hansen, A. Arvidsson, F. Å. Nielsen, E. Colleoni, and M. Etter, “Good friends, bad news-affect and virality in twitter,” *Future information technology*, pp. 34–43, 2011.
- [144] C. Budak, S. Goel, and J. M. Rao, “Fair and balanced? quantifying media bias through crowdsourced content analysis,” *Public Opinion Quarterly*, vol. 80, no. S1, pp. 250–271, 2016.

- [145] K. Leetaru, “Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space,” *First Monday*, vol. 16, no. 9, 2011.
- [146] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, “Bad is stronger than good.” *Review of general psychology*, vol. 5, no. 4, p. 323, 2001.
- [147] W. M. Johnston and G. C. Davey, “The psychological impact of negative tv news bulletins: The catastrophizing of personal worries,” *British Journal of Psychology*, vol. 88, no. 1, pp. 85–91, 1997.
- [148] P. F. Wu, “In search of negativity bias: An empirical study of perceived helpfulness of online reviews,” *Psychology & Marketing*, vol. 30, no. 11, pp. 971–984, 2013.
- [149] M. Trussler and S. Soroka, “Consumer demand for cynical and negative news frames,” *The International Journal of Press/Politics*, pp. 360–379, 2014.
- [150] D. S. Moore and S. Kirkland, *The basic practice of statistics*. WH Freeman New York, 2007, vol. 2.
- [151] Z. Yin, L. Cao, Q. Gu, and J. Han, “Latent community topic analysis: Integration of community discovery with topic modeling,” *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, p. 63, 2012.
- [152] Y. W. Teh, D. Newman, and M. Welling, “A collapsed variational bayesian inference algorithm for latent dirichlet allocation,” in *NIPS*, vol. 6, 2006, pp. 1378–1385.
- [153] M. J. Paul and M. Dredze, “Discovering health topics in social media using topic models,” *PLOS ONE*, vol. 9, no. 8, pp. 1–11, 2014.

- [154] K. W. Lim, C. Chen, and W. Buntine, “Twitter-network topic model: A full bayesian treatment for social network and text modeling,” in *NIPS2013 Topic Model workshop*, 2013, pp. 1–5.
- [155] S. Cucerzan, “Large-scale named entity disambiguation based on wikipedia data,” *EMNLP-CoNLL 2007*, pp. 708–716, 2007.
- [156] A. Islam and D. Inkpen, “Semantic text similarity using corpus-based word similarity and string similarity,” *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1–25, 2008.
- [157] Y. Koren and R. Bell, “Advances in collaborative filtering,” in *Recommender systems handbook*. Springer, 2015, pp. 77–118.
- [158] Z. Liu, X. Chen, Y. Zheng, and M. Sun, “Automatic keyphrase extraction by bridging vocabulary gap,” in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. ACL, 2011, pp. 135–144.
- [159] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, “Topics in semantic representation.” *Psychological review*, vol. 114, no. 2, p. 211, 2007.
- [160] R. Agrawal, R. Srikant *et al.*, “Fast algorithms for mining association rules,” in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.
- [161] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.
- [162] B. Goethals and M. J. Zaki, “Advances in frequent itemset mining implementations: report on fimi’03,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 109–117, 2004.

- [163] C. Borgelt, “An implementation of the fp-growth algorithm,” in *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*. ACM, 2005, pp. 1–5.
- [164] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, “Arnetminer: extraction and mining of academic social networks,” in *SIGKDD*. ACM, 2008, pp. 990–998.
- [165] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.
- [166] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *SIGKDD*. ACM, 2016, pp. 855–864.
- [167] R. Trivedi, M. Farajtbabar, P. Biswal, and H. Zha, “Representation learning over dynamic graphs,” *arXiv preprint arXiv:1803.04051*, 2018.
- [168] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, “Arima models to predict next-day electricity prices,” *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1014–1020, 2003.
- [169] F. Karim, S. Majumdar, H. Darabi, and S. Chen, “Lstm fully convolutional networks for time series classification,” *IEEE Access*, vol. 6, pp. 1662–1669, 2018.
- [170] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, “An lstm approach to short text sentiment classification with word embeddings,” in *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing*, 2018, pp. 214–223.

# List of Publications

## Journals:

- **Nagendra Kumar**, Gopi Ande, Jessu Shirish Kumar, and Manish Singh, *Toward maximizing the visibility of content in social media brand pages: a temporal analysis*, In *Social Network Analysis and Mining (SNAM)*, Vol. 8, p. 11, Springer, 2018.
- **Nagendra Kumar**, Rakshita Nagalla, Tanya Marwah, and Manish Singh, *Sentiment dynamics in social media news channels*, In *Online Social Networks and Media (OSNEM)*, Vol. 8, pp. 42 - 54, Elsevier, 2018.
- Anand Konjengbam, Neelesh Dewangan, **Nagendra Kumar**, and Manish Singh, *Aspect ontology based review exploration*, In *Electronic Commerce Research and Applications (ECRA)*, Vol. 30, pp. 62 - 71, Elsevier, 2018.

## Conferences:

- **Nagendra Kumar**, Yash Chandarana, Konjengbam Anand, and Manish Singh, *Using social media for word-of-mouth marketing*, In *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, pp. 391 - 406, Springer, 2017.
- **Nagendra Kumar**, Rahul Utkoor, Bharath Kumar Reddy Appareddy, and Manish Singh, *Generating topics of interest for research communities*, In *Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA)*, pp. 488 - 501, Springer, 2017.

- **Nagendra Kumar**, Anusha Yadandla, Suryamukhi K., Ranabothu Neha, Sravani Boya, and Manish Singh, *Arousal prediction of news articles in social media*, In *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE)*, pp. 308 - 319, Springer, 2018.
- **Nagendra Kumar**, Srikanth G, Karthik Yadav Mudda, Gayam Trishal, Anand Konjengbam, and Manish Singh, *Where to Post: Routing Questions to Right Community in Community Question Answering Systems*, In *Proceedings of Joint International Conference on Data Science and Management of Data (CoDS-COMAD)*, p. 7, ACM, 2019.
- Anand Konjengbam, Subrata Ghosh, **Nagendra Kumar**, and Manish Singh, *Debate Stance Classification using Word Embeddings*, In *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, pp. 382 - 395, Springer, 2018.
- Akilesh B., **Nagendra Kumar**, Bharath Reddy, and Manish Singh, *TRAFAN: Road traffic analysis using social media web pages*, In *Proceedings of the International Conference on Communication Systems & Networks (COMSNETS)*, pp. 655 - 659, IEEE, 2018.