Social Web Mining based on Link and Content Analysis

PhD Thesis

By
Hikmat Ullah Khan
55-FBAS/PHDCS/F09

Supervised By
Dr. Ali Daud

Department of Computer Science & Software Engineering
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad, Pakistan
2016
Praise be to Allah Almighty, Who Bestowed us life, health, family and His limitless Blessings. Peace and Blessings may Allah Almighty Bestow on our beloved Holy Prophet Hazrat Muhammad (PBUH) who is sent as a Mercy for the entire universe.

Dr. Ali Daud is highly dedicated for research achievements and is always available for research discussion. He is a highly committed researcher and aims high for his research achievements. I am also thankful to Prof. Dr. Muhammad Sher, Dean, Faculty of Basic and Applied Sciences and Dr. Husnain Naqvi, Chairman, Department of Computer Science and Software Engineering, for their consistent guidance.

The financial assistance provided from Higher Education Commission (HEC) to pursue my Ph.D research is duly acknowledged. This research work would not be completed without the generous help of HEC through Indigenous 5000 PhD fellowship.

I pay gratitude to foreign experts as well. Dr. Leonidas Akritidas, a post-doctoral researcher at University of Thessaloniki, provided me valuable comments and shared me how to write research articles. Thanks to Cyberemotions teams, especially Prof. Dr. Mike Thelwall and Dr. Georgios Paltoglou for providing datasets and replying all my queries.

I would like to extend thanks to my friends for their help: Mr. Sajjad Sadiq for help in implementation, Mr. Waqas Ahmed for suggestions in formal mathematical symbol notations, Dr. Khalid Khattak for comments about Algorithm representation, Syed Muhammad Saqlain, Dr. Muhammad Shoaib, Dr. Saqib Iqbal, Ch. Muhammad Shehzad Faisal and Tahir Afzal Malik for their valuable comments and support.

Last but not the least; I would like to thank all my beloved family members for their support and sparing me for the studies.


Declaration

“I hereby declare and affirm that this thesis neither as a whole, nor as part thereof has been copied out from any source. It is further declared that I have completed this thesis entirely on the basis of my personal effort, made under the sincere guidance of my supervisor. If any part of this report is proven to be copied or found to be a reproduction of some other, I shall stand by the consequences. No portion of the work presented in this report has been submitted in support of an application for other degree or qualification of this or any other university or institute of learning.”

Hikmat Ullah Khan

55-FBAS/PHDCS/F09
Dedication

I proudly dedicate my PhD dissertation to my parents:

My beloved father, Mr. Inayat Ullah Khan Niazi (Late), who always encouraged me and believed in me, but he is not here to see my achievements. I always miss him and pray that may Allah Almighty Shower His limitless Blessings on him.

My beloved mother, Mehr Fatima, who always prays for me and I pray that may Allah Give her a long, healthy and happy life. Amen.

Hikmat Ullah Khan
Abstract

The social web has emerged as a popular medium for providing and seeking information, views, opinions and social interaction. The social web sites such as online forums, web blogs, and wikis have a huge amount of user generated data in their archives. Analysing this content has important applications such as identifying influential users in online communities, improving information search in social platforms, and understanding users' behavior. Opinion mining is an active research area due to the availability of world-wide public views in the social web platforms. In this dissertation, three research problems in the social web are addressed: Subjectivity analysis of online forums using lexical and non-lexical features, finding the top influential bloggers and mixed-opinion classification of posts and threads in online forums using thread-specific feature set.

Opinion mining has vast applications such as online marketing, product evaluation and analysis of human nature. Subjectivity classification separates the content into subjective and non-subjective. Existing research efforts classify the text into subjective and non-subjective using dictionary or lexicon. In this research work, we propose the content-based dialog feature set for subjectivity classification of the posts. The proposed non-lexical features outperform the conventional lexical features. In addition, we propose and validate the significance of novel thread-specific feature set to classify the subjective and non-subjective threads. Experiments using supervised learning algorithms on the web forum dataset confirm that non-lexical features outperform conventional lexicon-based features for subjectivity classification. The role of each feature in the feature set is also discussed. The proposed dialog and thread-structure features do not need a lexicon and are applicable to unstructured content.

Finding the influential bloggers is an important research problem as it helps in e-commerce, online marketing and product knowledge searching. To identify the top influential bloggers, the existing methods consider basic features, but lack consideration of important features. A novel modular approach presents MIIB (Metric for Identification of Influential Bloggers) by introducing factors in terms of features related to bloggers’
productivity and popularity and the ranks of the blog sites. MIIB is compared with standard metrics and models and the analysis using the real-world blog dataset prove that the proposed metric is capable to identify the top influential bloggers in the Blogosphere effectively.

Thirdly, we classify mixed-opinion posts and threads in online discussions using content and link features. The mixed-opinion content contains both positive and negative emotions. Identification of the mixed-opinion content has potential applications such as making decisions about products, monitoring community views, and predicting public behavior. We present content-based post feature sets categorized into sentiment and dialog. At post level, dialog-act feature set outperforms lexicon-based sentiment feature set. We also introduce feature sets of sentiment, dialog and thread-specific categories for thread classification. The analysis of a dataset of public forum using several performance evaluation measures verifies that the proposed thread-specific features outperform sentiment and dialog feature sets. The role of each feature in the feature sets is debated. Our contribution is the proposal and validation of novel dialog and thread-specific feature sets which do not need a lexicon and are applicable to content in any language.
Research Achievements and Contributions

Journal Publications


Conference Publications


# Table of Contents

Chapter 1: Introduction............................................................................................................... 1
  1.1 Introduction.................................................................................................................. 2
  1.2 Social Web................................................................................................................ 2
      1.2.1 Web log .............................................................................................................. 4
      1.2.2 Web Forum ..................................................................................................... 4
  1.3 Investigated Research Topics ..................................................................................... 5
      1.3.1 Identification of Top Influential Bloggers ....................................................... 5
      1.3.2 Subjectivity Analysis using Non-Lexical Features .......................................... 6
      1.3.3 Mixed-Opinion Classification using Thread Features .................................... 7
  1.4 Existing Issues and Research Challenges .................................................................... 7
      1.4.1 Issues in Social Web Content .......................................................................... 8
      1.4.2 Issues in Finding Influential Bloggers ............................................................. 8
      1.4.3 Issues in the Use of Sentiment Lexicons ........................................................ 9
  1.5 Research Contributions ............................................................................................. 10
  1.6 Dissertation Outline .................................................................................................. 11

Chapter 2: Related Work ........................................................................................................ 12
  2.1 Subjectivity Analysis ................................................................................................... 13
      2.1.1 Opinion Mining ............................................................................................... 13
      2.1.2 Approaches for Opinion Mining ...................................................................... 14
      2.1.3 Subjectivity Classification ............................................................................... 15
  2.2 Finding Top Influential Bloggers ................................................................................ 18
  2.3 Mixed-Opinion Classification of Web Forums ........................................................... 22
      2.3.1 Mixed-Opinion Content Classification .............................................................. 22
      2.3.2 Feature-Centric Classification of Online Forums .............................................. 23

Chapter 3: Subjectivity Analysis using Non-Lexical Features .............................................. 26
  3.1 Introduction ................................................................................................................ 27
  3.2 Why Subjectivity Analysis of Online Forums ............................................................ 27
  3.3 Problem Formulation and Problem Statement .......................................................... 28
      3.3.1 Problem Formulation ...................................................................................... 28
      3.3.2 Problem Statement ........................................................................................ 29
Chapter 5: Social Web Mining based on Link and Content Analysis

5.1 Introduction
5.2 Why Classifying Content for Mixed-Opinion Identification
5.3 Problem Formulation and Problem Statement
   5.3.1 Problem Formulation
   5.3.2 Problem Statement

Chapter 4: Identification of Influential Bloggers

4.1 Introduction
4.2 Why Finding Influential Bloggers
4.3 Problem Formulation and Problem Statement
4.4 MIIB: The Proposed Metric
4.5 Experimental Setup
   4.5.1 TUAW Dataset
   4.5.2 Performance Evaluation Measures
4.6 Results and Discussion
4.7 MIIB Evaluation using Engadget and Techcrunch Dataset
   4.7.1 Engadget Dataset
   4.7.2 Comparison of MIIB using Engadget with Baseline metrics
   4.7.3 Techcrunch Dataset
   4.7.4 Comparison of MIIB using Techcrunch data with existing metrics
4.8 Chapter Summary

Chapter 3: Mixed-Opinion Classification using Thread Features

3.4 Features Engineering
   3.4.1 Post Features
   3.4.2 Thread Features
3.5 Experimental Setup
   3.5.1 Classification Algorithms
   3.5.2 Data
   3.5.3 Performance Evaluation Measures
3.6 Results and Discussion
   3.6.1 Post Results
   3.6.2 Thread Results
3.7 Most Useful Features
3.8 Chapter Summary

Chapter 2: MIIB: The Proposed Metric

2.1 Introduction
2.2 Why Finding Influential Bloggers
2.3 Problem Formulation and Problem Statement
2.4 MIIB: The Proposed Metric
2.5 Experimental Setup
   2.5.1 TUAW Dataset
   2.5.2 Performance Evaluation Measures
2.6 Results and Discussion
2.7 MIIB Evaluation using Engadget and Techcrunch Dataset
   2.7.1 Engadget Dataset
   2.7.2 Comparison of MIIB using Engadget with Baseline metrics
   2.7.3 Techcrunch Dataset
   2.7.4 Comparison of MIIB using Techcrunch data with existing metrics
2.8 Chapter Summary
5.4 Feature Engineering ................................................................. 67
  5.4.1 Post Features ........................................................................ 68
  5.4.2 Thread Features ................................................................. 70
5.5 Experimental Setup ............................................................... 74
  5.5.1 Classification Algorithms .................................................. 74
  5.5.2 Dataset ............................................................................. 75
  5.5.3 Performance Evaluation Measures ..................................... 75
5.6 Results and Discussion ........................................................... 77
  5.6.1 Post Results ....................................................................... 77
  5.6.2 Thread Results ................................................................. 78
  5.6.3 Comparative Performance of Classification Algorithms ........ 79
  5.6.4 Relative Performance of Features ...................................... 79
  5.6.5 Result analysis using ROC, AUC, Lift and Cost .................. 80
  5.6.6 Most Useful Features ........................................................ 81
5.7 Chapter Summary ................................................................. 82

Chapter 6: Summary, Conclusion and Future Directions .................. 84
  6.1 Summary ............................................................................. 85
  6.2 Conclusions ....................................................................... 86
  6.3 Future Directions ............................................................... 87
References .................................................................................. 88
List of Tables

Table 1: A Comparison of Formal Content vs Social Web Content ........................................... 8
Table 2: An overview of Existing Work in Subjectivity Analysis ............................................ 17
Table 3: An overview of Existing Feature based Models .......................................................... 21
Table 4: An Overview of Existing work for Mixed-Opinion Content Classification ................. 24
Table 5: Sample Sentences/topics differentiating Subjective vs Non-Subjective ...................... 27
Table 6: List of Symbols used for Subjectivity Analysis .............................................................. 29
Table 7: List of Proposed Post Features for Subjectivity Analysis ........................................... 32
Table 8: List of Baseline Thread Features for Subjectivity Analysis ....................................... 33
Table 9: List of Proposed Thread Features for Subjectivity Analysis ...................................... 33
Table 10: A Comparison of Dataset Characteristics .................................................................... 38
Table 11: Feature Set-wise Post Results for subjectivity analysis ........................................... 40
Table 12: Thread Classification Results for Subjectivity Analysis ........................................... 41
Table 13: Top Post Features for Subjectivity Analysis .............................................................. 42
Table 14: Top Thread Features for Subjectivity Analysis ......................................................... 43
Table 15: List of Symbols used for the Proposed Metric ............................................................ 47
Table 16: Features’ List and their Objectives .............................................................................. 48
Table 17: Statistics of the TUAW Dataset .................................................................................. 52
Table 18: The top Ten Bloggers using Individual Feature ......................................................... 55
Table 19: Top Ten bloggers based on Modules, iFinder and the MIIB .................................... 56
Table 20: Top Ten bloggers based on Modules, iFinder and the MIIB .................................... 57
Table 21: Evaluation of MIIB vs Baseline Metrics ..................................................................... 57
Table 22: Evaluation of MIIB and its Modules using OSim ....................................................... 59
Table 23: Evaluation of MIIB and its Modules using Spearman Rank-order Correlation ......... 59
Table 24: Evaluation of MIIB and its Modules using Kendall Correlation ............................ 59
Table 25: Engadget Dataset Characteristics ............................................................................ 60
Table 26: A Comparison of MIIB, Modules versus Baselines ................................................. 61
Table 27: A comparison between MIBSA and baseline metrics ............................................ 62
Table 28: Techcrunch Dataset Characteristics ......................................................................... 63
Table 29: Blogger's Ranking based on MIIB, Modules and Baseline ..................................... 63
Table 30: List of Symbols used for Mixed-Opinion Classification .......................................... 68
Table 31: List of Post Features for Mixed-Opinion Classification .......................................... 69
Table 32: List of Thread Features for Mixed-Opinion Classification .................................... 71
Table 33: BBC Forum Dataset Statistics .................................................................................. 75
Table 34: Post Results for Mixed-Opinion Classification ....................................................... 77
Table 35: Thread Results for Mixed-Opinion Classification ................................................... 78
Table 36: Results to Find Mixed-opinion Threads using Feature Set Combination ................ 78
Table 37: Overall Comparison of the Feature Sets .................................................................. 79
Table 38: Comparative analysis using ROC Accuracy, AUC, Lift and Cost Results ............ 81
Table 39: Ranking of Top Post Features .................................................................................... 81
Table 40: Ranking of the Top Thread features ......................................................................... 82
Table of Figures

Figure 1: An overview of Activities on Social Web............................................................... 3
Figure 2: Layout of the Proposed Framework for Subjectivity Analysis ......................... 32
Figure 3: Power Law Analysis of BBC dataset using Log Scale Chart.......................... 38
Figure 4: Post Classification Results for each Feature set .............................................. 40
Figure 5: Thread Results for Subjectivity Analysis ......................................................... 41
Figure 6: Layout of the Proposed Framework for MIIB ............................................... 49
Figure 7: The order-alterations of the Influential Bloggers using individual Features .. 55
Figure 8: A Module-wise Comparative Analysis ......................................................... 58
Figure 9: Comparison of MIIB and its Modules ............................................................... 58
Figure 10: The proposed framework for mixed-opinion classification of Online Forum .. 74
Figure 11: Classification Results using each Feature Sets ............................................. 80
Figure 12: Classification Results using Feature Set Combinations ................................ 80
# Table of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>Agglomerative Double Clustering</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>ME</td>
<td>Maximum Entropy</td>
</tr>
<tr>
<td>MEIBI</td>
<td>“MEtric for Identifying a Blogger’s Influence”</td>
</tr>
<tr>
<td>MEIBIX</td>
<td>“MEtric for Identifying a Blogger’s Influence eXtended”</td>
</tr>
<tr>
<td>MIIB</td>
<td>Metric for Identification of Influential Bloggers</td>
</tr>
<tr>
<td>MDL</td>
<td>Maximum Description Length</td>
</tr>
<tr>
<td>MPQA</td>
<td>Multi Perspective Question Answering</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>ODM</td>
<td>Oracle Data Miner</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristics</td>
</tr>
<tr>
<td>SIIB</td>
<td>Semantically Identified Influential Bloggers</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TREC</td>
<td>Text Retrieval and Evaluation Conference</td>
</tr>
<tr>
<td>TUAW</td>
<td>The Unofficial Apple Weblog</td>
</tr>
<tr>
<td>URL</td>
<td>Universal Resource Locator</td>
</tr>
<tr>
<td>WWW</td>
<td>World Wide Web</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction
Chapter 1

Introduction

1.1 Introduction

This chapter presents an introduction of the social web and its two important platforms of web blogs and forums. We introduce our three research topics addressed in this thesis: 1) Finding top influential bloggers in the blogging community, 2) the subjectivity analysis of online forums using conventional lexical and proposed non-lexical features, 3) Mixed-opinion classification of posts and threads in online forums using novel thread-structure features. Research issues and challenges in the social web are explored. The main research contributions are briefly listed before presenting the thesis outline.

1.2 Social Web

The Social Web, or Web 2.0 [1] is a set of web sites and tools that enable people to generate their own content on the World Wide Web (WWW). The social web platforms like web blogs, online forums, social networks and media sharing sites provide facility to do have social interactions and create social activities. The users in such platforms create virtual communities also referred as the social networks which let users share their comments, opinions, experiences and knowledge with other users.

The social web presents new and unique opportunities and challenges. It is a source of shifting the focus from classical issues of engineering and computer science to natural science approach. Unlike the engineering system, where the focus is on controls, the social web researchers have started to focus on the real world issues by studying users’ activities and their social interactions. This acts as a bridge between computer science, and other sciences such as social science, economics, psychology and the nature of complex systems.
Introduction

The latest statistics\textsuperscript{1} show the exponential growth of the social web as the internet users have increased from the mere 1\% to about 45\% of the world population since 1995. There are about 3 billion active internet users of which over 2.1 billion are social web users. Over 3.65 billion cell phone users enjoy internet access and half of them are active users of social media. Figure 1 shows main social web platforms depicting the diversity of the online activities.

Figure 1: An overview of Activities on Social Web\textsuperscript{2}

There are various social web platforms, but in this work, we focus on web blogs and forums which are briefly discussed.

1.2.1 Web log
A web log (blog as a short) is a social web channel consisting of posts usually displayed in reverse chronological order. Blogs are classified into two types: the community weblogs and the individual weblogs. An individual blog is usually related to a focused topic as its posts are created by single blogger and other users can comment on it. Whereas, a community blog covers diverse topics as all the blog members may initiate and comments a blog post. In a blog, users may generate any content by writing own text, sharing photos and videos, and linking to other posts and sites. The blogosphere is the universe of all the web blogs which includes millions of active bloggers and billion readers. According to Wikipedia³, blogosphere consist of above 172 million blog sites producing about million blog posts every day.

Blog are significant as the others follow expert and celebrities for their views, opinions, information. Blog users share their activities and discuss with each other their experiences by expressing their views and opinions. As in the real world, we consult others before taking a decision such as buying any product, attending an event or planning for a tour. Now, people discuss and seek opinions of expert and ask for experiences from other users who are influential. Thus, finding such top bloggers who can have an impact on others is a vital research question.

1.2.2 Web Forum
An online forum usually consists of several sub-forums covering various topic categories. Each sub-forum further consists of numerous threads. A thread is a focused top-specific discussion containing a number of posts. A post is a comment created by a community member. It may consist of textual content, outlinks to other posts in the forum or web sites and multimedia content such as images, videos. A post presents public views, reviews and a thread presents a public discussion among common users. This feature motivates researchers to analyze and mine its latent characteristics.

An online forum has unique structural and textual features that distinguish it from a conventional web page. An online forum consists of a number of threads and each thread containing multiple posts. A user starts a new thread by creating first post which sets up the topic for comments and discussion. The content in the forum can be of different nature and can have different text, images, and outgoing hyperlinks to other web pages and threads in the forums. The social web researchers explore the thread content and its unique characteristics for study of various research problems.

1.3 Investigated Research Topics

In this thesis, we focus our research on finding top bloggers, subjectivity analysis and identification of mixed-opinion content in online forums.

1.3.1 Identification of Top Influential Bloggers

Blog users express their views, opinions and experiences regarding various topics. The facility of social interaction motivates researchers to add social concepts in their research about understanding the human nature and behavior indirectly. A study presents that, in real life, 83% of people consult others, especially their family fellows, friends or trustworthy persons instead of trusting advertisements before planning for a dinner to a hotel, 71% of them do the same ahead of visiting a place or purchasing a new product, and about 61% of people act likewise before watching a new movie. It is a common practice that while taking any decision, people seek others and listen to their views, experiences, and recommendations. So, the individuals whose and recommendations and views have impact on others are known as, in the relevant literature, the influential [2].

The exponential growth of the web 2 motivates scholars to address issues related to the blogosphere [3]. An early work finds influential blogs [4]. Ranking algorithms such as PageRank [5] are used for ranking authors in scholarly system [6], but the blog graph sparseness renders ranking approaches inappropriate for blogosphere research. Later, a number of metrics and models are proposed to identify the top bloggers who influences other users in the blogging community which are discussed in chapter 2 in detail.

This research proposed MIIB (“Metric for Identification of Influential Bloggers”) by introducing factors. A novel modular approach introduces three modules which are
Introduction

named as Popularity, Productivity, and BlogRank. Productivity measures the productive nature of a blogger, popularity computes a blogger’s influence with community and BlogRank depicts a blog’s significance where the blogger initiates his/her posts. The evaluation using standard performance evaluation measures metric against the standard existing models and metrics using the real world dataset confirms that the proposed metric identifies influential bloggers in an effective manner.

1.3.2 Subjectivity Analysis using Non-Lexical Features

Opinions and sentiments are part of public discussion and the social web forum discussion is not an exception. Opinion mining is an active research area due to participation of people from all over the world who discuss about various topics related to all spheres of our daily life and it has vast applications such as analyzing the products, checking market trends, studying social issues and examining human behavior. Sentiment mining addresses research problems including sentiment extraction, document summarization, subjectivity and sentiment classification.

Subjectivity analysis deals to separate factual information and opinionated information. It classifies a document as subjective or non-subjective/objective. An objective document presents factual and informative content. Such content is usually related to a technical topic or provides real information about a place, event or any other topic. On the other hand, a subjective document or content provides public points of views, opinions and emotions. Subjectivity classification has vast applications in various fields such as to find the opinions of customers in online review sites, to improve the quality of answers of opinion questions in question-answering forum, to summarize multi-documents, etc. The existing works in opinion and subjectivity mining use dictionary and lexicon for detection of sentiment [7]. Various supervised learning methods have also been applied for subjectivity classification and opinion mining [8].

In this work, we first propose content based dialog feature sets at post level and validate that the proposed features outperform conventional lexicon based approach to classify posts into factual or opinionative categories. Then, we propose thread-specific features that also need no dictionary or lexicon and are helpful to differentiate subjective threads
Introduction

1.3.3 Mixed-Opinion Classification using Thread Features

In the social web forums, users express their views and opinions from anywhere in the world and can discuss any topic of interest. A large number of users discuss social, religious, political and technical issues. An essential aspect of public discussions is sentiment and usually all such topics have heated conversations where people share their arguments in favor of their point of view or to disagree with others’ views. Opinion mining analyzes the users’ opinions and emotions towards different products, issues and people [9]. Various opinion mining research studies analyze the social web content and help us to have a deep understanding of the human behavior [10-13]. The subjectivity classification classifies the content into subjective or non-subjective [14] and sentiment classification categorizes the opinions into positive or negative. Identification of mixed opinions, having positive emotions and negative emotions is one of the domains of opinion mining.

Our aim is to find such posts and discussions in online forums where users express high mixed opinions i.e., having both high positive and high negative emotions. We consider lexicon-based sentiment and non-lexical dialog act feature sets to find opinion posts. For threads, similar sentiment and dialog feature sets are used and thread-specific features are introduced which exploit the structure of the threads. The proposed link-based feature set outperforms the conventional lexicon-based sentiment and content-based dialog features to find high mixed-opinion threads. This feature-centric approach uses a large real world dataset of online forum containing threads of various topics. The thread-structure features which do not need a lexicon and are computable irrespective of the content language.

1.4 Existing Issues and Research Challenges

Let us here discuss main issues faced while analyzing the social web content. The freestyle nature of the social web content is discussed which poses research challenges while analyzing its content. The issues about finding the influential bloggers in the community are shared. In addition, various issues are explored about the use of lexicons.
Introduction

for opinion analysis to confirm that there is a requirement of new methods which does not need a lexicon and can still be useful for opinion mining.

1.4.1 Issues in Social Web Content

In social web channels, users are free to generate any content in any form using various languages and can add various symbols, emoticons. Users comments are usually grammatically incorrect due to an informal writing styles, spelling mistakes, abbreviations, hashtags, etc. [15]. User generated content is difficult to be classified as ironic or cynical. Thus, this short-coming may conclude towards an improper opinion mining and finding an erroneous orientation [15, 16]. The free content in messages or online forums, known as texting, contains a number of slang and abbreviated words. Online web content is multilingual and has such nonstandard words which cannot be recognized by any lexical database. The table 1 illustrates the difference between the formal content and free style which depicts that such freestyle informal social web content cannot be analyzed using any standard dictionary or lexicon.

<table>
<thead>
<tr>
<th>Informal Web content</th>
<th>Formal Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>btw wenz ur flt 2moro</td>
<td><em>By the way, when is your flight tomorrow?</em></td>
</tr>
<tr>
<td>Hi, hru, im f9, wht r u dng</td>
<td><em>Hi, How are you, I am fine, what are you doing?</em></td>
</tr>
<tr>
<td>Ic. That’s gr8. It’s getting l8. V shd go b4 it gts dark. cu l8r.</td>
<td><em>I see that is great. It is getting late. We should go before it gets dark. See you later.</em></td>
</tr>
</tbody>
</table>

1.4.2 Issues in Finding Influential Bloggers

The research domain of finding influential bloggers is subjected to certain issues. The main challenge is to define the bloggers’ influence as the idea or concept of influence is subjective, thus it is tough to quantify into a metric or a model. Therefore, we find limited relevant research work as illustrated in chapter 2. Another open discussion in this research domain is to differentiate influential bloggers from active ones. One of the main issues is the lack of ground truth and standard benchmark datasets. The problem cannot

---

4 We find many internet slang dictionary, one large and well known is as follows: Slang dictionary – Text Slang & Internet Slang Words. [http://www.noslang.com/dictionary/](http://www.noslang.com/dictionary/)
be marked as a machine learning technique. The researchers solely depend on certain features and the use of statistics for computation of those features. But such approach raises questions on models and metrics evaluation due to absence of ground truth [17]. Therefore, the existing works evaluate and compare the results using feature-centric approach. The absence of ground truth lacks the use of performance evaluation measures as well.

1.4.3 Issues in the Use of Sentiment Lexicons

Bing Liu [18] shares that sentiment words are significant for opinion analysis and he also argues that using sentiment considering sentiment words identifying using sentiment lexicon is not enough. Main issues are as follows:

1. A word may likely have opposite orientation for a different field or in another domain. For instance, “suck” shows negativity usually such as the sentence “this device sucks”, but it may also give some positive opinion, for instance, “This new vacuum clear really sucks”.

2. A sentence may consist of emotional words, but, in real sense, may not depict any opinion. Question and conditional sentences are typical examples, such as, “If I may find a good camera, I will surely buy it” have positive words of “help”, “good”, “surely”, but, in real sense, neither shows a positive or negative sentiment about a particular product, in this case, a camera.

3. Sarcastic sentences are also difficult to analyze even having sentiment word or not. e.g., “What a great camera! It stopped functioning in a week”. Such words are common in public discussions about topics such as products, politics and social issues.

4. A sentence may imply an opinion without the use of sentiment word, e.g., “This camera uses a lot of battery power” implies negative emotion regarding a camera but actually it provides factual information and does not contain any negative word.

There is a need to bridge the gap between unstructured multimodal information and structured machine process able data [19]. To cater to above mentioned issues, we
propose dialog act and thread-structure feature sets which do not need a lexicon for computation.

1.5 Research Contributions

We address three research problems from the social web research domain. At first, existing works focus on subjectivity analysis using lexical features and syntactic features like POS (Parts of Speech)-tags, n-gram, subjectivity clues. Our main contribution is as below:

1. Proposal of non-lexical feature set for subjectivity analysis of post and threads of web forums: content based dialog features and thread-structure features and verification that the new features help to classify subjective content in online forums and identification of the role of various features.

In the second, existing research focused on proposing models and metric to find influential bloggers in the blogging community, but lack to consider important features, provide linear approach, our research contribution in this regard are as follows:

1. Design and proposal of a metric using modular approach to find influential bloggers by introducing features with weights and then evaluating the metric against the standard existing model and metrics using feature-based and performance evaluation measures confirming that the proposed metric identifies the influential bloggers effectively.

In the third, existing research works aim to find high mixed-opinion content using natural language processing approaches and mainly using lexicons to find the positive and negative sentiments in documents. This approach is unsuitable for new multilingual and content not suitable for finding high sentiments from posts’ content. Our contribution includes:

2. Proposal of several content-based dialog features and thread-specific features and validation by application of classification techniques to confirm that the novel features, not needing a lexicon, outperform lexical features to identify mixed-opinion posts and threads in online forums and identification of the role of various features.
1.6 Dissertation Outline

The dissertation is divided into six chapters, starting from this introductory Chapter 1, which provides an introduction to the social web and explores research issues and shares research contribution. We introduce three research domains of influential bloggers identification, subjectivity analysis and opinion-mixture classification. Chapter 2 reviews the related work about the three modules of the dissertation. The three proposed research work have been described chapter-wise from chapter 3 to 5. The chapters 3, 4 and 5 each completely cover the addressed research problem formulation, proposed framework, feature engineering, proposed algorithms, results evaluation and discussion. Chapter 6 provides conclusions and potential future directions. Next to chapter 6, a list of research works is given which are referred in this thesis.
Chapter 2: Related Work

“Social Web Mining based on Link and Content Analysis”
Chapter 2
Related Work

This chapter presents the existing related work in the research problems investigated in this research thesis: Finding influential bloggers, Subjectivity analysis and the Mixed-Opinion Classification of online forums.

2.1 Subjectivity Analysis
The related work reviews the earlier recent work related to opinion mining, subjectivity classification and online forums.

2.1.1 Opinion Mining
Opinion mining identifies the expressed opinion on a certain topic and evaluates its polarity (either positive or negative). It targets to extract and process users’ opinion about various topics. Opinion mining includes subjectivity classification (to distinct objective and subjective text) and opinion/sentiment classification (to recognize the positive, neutral or negative polarity of the opinions [8, 19]. Due to user generation ability, opinion mining also finds opinion in content produced in the social web channels for various reasons like to summarize the product reviews [20] and find opinions in online review sites [21]. An online review has a single role as it depicts user’s viewpoint about a certain topic or a product. Whereas, a thread structure comprises of a number of posts possibly from various users and such posts serve multiple roles like question, comment, feedback, junk, etc. [22].

Opinion mining uses diverse feature sets for finding opinions in the social web. A feature-based sentiment classification approach introduces for finding opinion documents in web sites [23]. Another model uses cognitively inspired frames for the detection of holder, topic and sub-topic of opinion [24]. The lexical features find sentiment in the product review site [21]. Feature-centric approaches finds opinions in web documents [23]. Twitter messages have also been analysed to identify political opinions to predict
political preferences [25], analyse stock market [26], and perform statistical analysis of linguistic features [27]. Another work also identifies various differences between opinion mining in microblogs and conventional sentiment analysis of documents [27].

2.1.2 Approaches for Opinion Mining

There are various methods for opinion finding and sentiment analysis, but mainly sentiment lexicons and machine learning techniques have been used. Here, let us briefly describe both main methods which are also used in our work as well.

2.1.2.1 Sentiment Lexicon

The obvious indicator of sentiment is opinion word (or sentiment words) which express either positive or negative sentiment. As an example, nice, pleasing, and marvellous are the positive words, whereas ugly, pathetic, poor, ugly, and dreadful are the negative words. A collection of such words and phrases that expresses opinion is known as opinion lexicon. A number of lexicons have been developed by the researcher over the years. We find a number of opinion lexicons the following well known lexicons including:

- Sentiment Lexicon [20]
- Multi-Perspective Question Answering (MPQA) lexicon [28]
- SentiWordNet [29] [30]
- Emotion Lexicon [31]

Over the years, the sentiment lexicons have vastly been developed and applied. We find related work that create lexicon for special target content. For instance, Blog06 corpus [32] is constructed for blog data and a subjective lexicon is created using Text Retrieval and Evaluation Conference (TREC) datasets [33]. Multilingual lexicons have also been developed, such as NTCIR lexicon which is developed for multilingual opinion analysis task (MOAT) [28]. In recent years, research works target to construct opinion lexicons for one particular language, such as for Arabic Language [34], Spanish lexicon [35], Chinese lexicon [36]. One of the most widely used lexicon is SentiWordNet [30] as it is an outcome of automatic annotation of the WordNet synsets with respect to their positive, negative and neutral degrees.
The lexicon based approach has advantage that it is easy and quickly finds words with their respective sentiment orientation. But the main drawback is that sentiment orientation of words is general that makes it hard to achieve domain specific orientation of words. Other limitations have already been discussed in detail in Section 1.4.3 of chapter 1.

2.1.2.2 Machine Learning Approaches

Sentiment classification is primarily a text classification problem and usually classifies a given document or text as positive or negative. Classification algorithms like Naïve Bayes algorithm (NB), Tree based technique such as Decision Tree (DT) and kernel based Support Vector Machine (SVM) are applied. First research work which applied supervised learning approach for text classification classified using SVM and NB methods [37].

Over the years, we find a number of feature-centric approaches for sentiment classification. Syntactic relations have been combined with traditional features [38]. The sentiment valence based on contextual and sentiment shifters have been used using semi-supervised learning techniques [39]. Instead of labelled document, labelled features have been used in [40]. A tree-based classification approach [41] has been proposed that uses the Conditional Random Field (CRF) technique. Personal and impersonal sentences were used as features to check their role in sentiment classification [42]. Linguistic and other features that identify the informal or creativity nature of the content have been exploited [43]. Classification algorithms has also been compared using different datasets [44]. SVM has been compared with Artificial Neural Network (ANN) for sentiment classification at document level [45]. Objective words have been identified using SentiWordNet to improve the sentiment classification in online reviews [46].

2.1.3 Subjectivity Classification

Subjectivity classification, a domain of opinion mining, classifies content into subjective and non-subjective. The comprehensive survey discusses the subjectivity as an emerging research domain [8]. An early work generates a gold standard for subjectivity classification [47] and classify sentences using features like the existence of an adjective, a pronoun or a modal in the sentences. Subjectivity classification is applied on WordPress
articles using an un-annotated data using classifiers to generate training data [48]. The subjectivity patterns are also learnt by applying semi-supervised learning approach. Word sense subjectivity classification has been applied on un-annotated data to prove that the proposed approach outperform earlier supervised training sets [49].

A recent survey discusses the various categories of feature sets for opinion mining [50]. It shares that lexical and linguistic features are main source of the features to find opinions in the text. It also covers the different types of opinion mining such as subjectivity analysis. It covers the various limitations of the lexical approach as well. Another recent review on subjectivity analysis also encompass various techniques and shares the pits and falls of various existing techniques [51]. It also verifies that non-lexical feature set has not yet been used. A recent work shares that due to problem in finding opinion in text, it is tough to classify subjectivity mining of multilingual content using lexicon based approaches [52]. The question of how to find subjective and opinionated content in the social web forums has also been discussed in a contemporary research work [53]. A recent survey of sentiment and subjectivity analysis provides overview of the natural language processing techniques applied for sentiment analysis and presents the advantages and disadvantages of such approaches [54].

Prakhar Biyani et al. [55] propose feature-centric approach to carry out subjectivity analysis and to find information type within forum threads. The features include linguistic features such as POS tags and their combinations. This research considers the initial post in the thread as its topic. Due to free content generation availability in online forum, we find multi-lingual content in public forums. Subjectivity classification in multiple languages was initiated by proposing a method to develop certain tools that use the machine translation facility and analyses the subjectivity of the content [56]. Subjectivity and opinion analysis for Arabic social media explore various differences in the target language as compared to conventional methods [57]. One of the recent work proposed framework for multilingual feature set for subjectivity analysis [58]. The overview of the existing works in this research domain is given in Table 2.
Table 2: An overview of Existing Work in Subjectivity Analysis

<table>
<thead>
<tr>
<th>Ref</th>
<th>Main Idea</th>
<th>Classification</th>
<th>Algo &amp; Features</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>[51]</td>
<td>Text classification</td>
<td>ML</td>
<td>SVM and NB and Max Entropy (Words Features)</td>
<td>Cornell Movie review dataset</td>
</tr>
<tr>
<td>[52]</td>
<td>Sentiment Classification using Syntactic relationships</td>
<td>ML</td>
<td>NB, SVM (Word related features)</td>
<td>Cornell Movie review dataset</td>
</tr>
<tr>
<td>[53]</td>
<td>Sentiment analysis using contextual and sentiment shifters</td>
<td>Semi-supervised</td>
<td>, SVM (Polarity shifter features)</td>
<td>Product Reviews</td>
</tr>
<tr>
<td>[54]</td>
<td>Identify political standpoint</td>
<td>Semi-supervised</td>
<td>(Linguistic &amp; opinionative Features)</td>
<td>Political data of project Vote Smart Website</td>
</tr>
<tr>
<td>[55]</td>
<td>Sentiment classification</td>
<td>Lexicon</td>
<td>Labelled features (lexicon-based features)</td>
<td>Movie review data and multi-domain sentiment dataset</td>
</tr>
<tr>
<td>[56]</td>
<td>Sentiment classification using dependency tree-based classification</td>
<td>CRF</td>
<td>(Bag-of-Features)</td>
<td>Four corpora: ACP corpus, NTCIR corpus, TE corpus NTT Blog corpus</td>
</tr>
<tr>
<td>[57]</td>
<td>Sentiment Classification</td>
<td>Supervised and Semi-supervised</td>
<td>(Personal and impersonal sentences features)</td>
<td>Product Reviews</td>
</tr>
<tr>
<td>[58]</td>
<td>Identify informal or Creativity nature of content</td>
<td>ML</td>
<td>SVM, NB, ME (Linguistic Features)</td>
<td>Twitter Dataset</td>
</tr>
<tr>
<td>[59]</td>
<td>Comparison of classification algorithms</td>
<td>Lexicons, ML</td>
<td>(Linguistic and Pattern Features)</td>
<td>Movie Review</td>
</tr>
<tr>
<td>[60]</td>
<td>Document-level Sentiment Classification</td>
<td>ML</td>
<td>SVM and ANN(Artificial Neural Network)</td>
<td>Movie Review dataset</td>
</tr>
<tr>
<td>[61]</td>
<td>Sentiment Classification of Online reviews</td>
<td>Lexicon</td>
<td>Lexical Features (SentiWordNet)</td>
<td>Web Review Forum</td>
</tr>
<tr>
<td>[63]</td>
<td>Separating subjective vs objective classifiers using un-annotated data</td>
<td>ML, Lexicon</td>
<td>(NB), Lexical features</td>
<td>Multi-Perspective Question Answering Dataset</td>
</tr>
<tr>
<td>[64]</td>
<td>Word sense subjectivity labelling</td>
<td>ML, Lexicon</td>
<td>Linguistic Features</td>
<td>Movie dataset and MPQA Dataset</td>
</tr>
<tr>
<td>[65]</td>
<td>Analyzing subjectivity in threads</td>
<td>Lexicon, ML</td>
<td>(Lexical features, limited sentiment thread features)</td>
<td>NYC and Ubuntu Forum</td>
</tr>
<tr>
<td>[66]</td>
<td>Generate resources for subjectivity annotations for new language</td>
<td>Lexicon</td>
<td>Lexical features</td>
<td>OpinionFinder Corpus</td>
</tr>
<tr>
<td>[67]</td>
<td>Subjective and sentiment analysis in Arabic social media</td>
<td>ML</td>
<td>SVM, Lexical Features</td>
<td>DARDASHA data</td>
</tr>
</tbody>
</table>
Prakhar Biyani et al. [55] propose feature-centric approach to carry out subjectivity analysis and to find information type within forum threads. The features include linguistic features such as POS tags and their combinations. This research considers the initial post in the thread as its topic. Due to free content generation availability in online forum, we find multi-lingual content in public forums. Subjectivity classification in multiple languages was initiated by proposing a method to develop certain tools that use the machine translation facility and analyses the subjectivity of the content [56]. Subjectivity and opinion analysis for Arabic social media explore various differences in the target language as compared to conventional methods [57]. One of the recent work proposed framework for multilingual feature set for subjectivity analysis [58]. The overview of the existing works in this research domain is given in Table 4.

2.2 Finding Top Influential Bloggers

As the Blogosphere in the social web is growing exponentially in last decade, the researchers’ focus shifted to work in this domain. The early focus has been to find the important web blogs in the blogosphere. The concept of identification of the influential blogs is discussed to show how the influence of the blogs should be measured [4]. The information dissemination in the blogosphere is studied [59] which help to understand how the topics are discussed and information is propagated. Later, the model to measure the spread of the influence of the information in the in the blogosphere is proposed [60]. The main purpose was to study the web blogs and the spread of influence. A comprehensive review discusses of the various research issues, tools and application about it [17].

The domain of influential bloggers identification is introduced by proposing a basic model, iIndex, which is also known as influence flow model as it considers the basic feature of inlink and comments and targets to measure the influence flow in the blogosphere [61]. This model considers the factors related to the bloggers and their posts and finds top influential bloggers in The Unofficial Apple Weblog (TUAW) dataset. It compares top influential bloggers with active bloggers. Another work that discusses the temporal aspect about the influence of a blog post and its bloggers and concludes that the
importance of a post decreases as the passes by [62]. This work finds top influential bloggers in the TUAW blog considering time as an important feature.

WisClus [63] identifies top bloggers from the individual blogs as well as community blogs using BlogCatalog dataset. It uses link-based clustering approach by creating a graph of relations in which category label has been considered as nodes and their link strength has been taken as edges. The model, Semantically Identified Influential Bloggers (SIIB), [64] identifies the top bloggers by adopting the features of iIndex and introducing the semantic features and compare results with quantitative and semantic measures. The authors of iIndex introduce a widely considered standard and comprehensive model, iFinder, based on features including activity generation (number of posts initiated), recognition (number of inlinks and comments received), novelty (outlinks) and eloquence (post-length) using TUAW dataset [65]. It considers limited but important feature set and finds the top influential bloggers. It compares influential bloggers with active bloggers. It lacks certain features such as a blogger’s consistency and the significance of the blogsites where the bloggers post their blogs. The evaluation is carried out against PageRank, which is widely used ranking algorithm, while it is also argued in the paper that ranking approaches are inappropriate for the research work related to blogosphere.

To find the influential bloggers, the metrics “Metric for Identifying a Blogger’s Influence” (MEIBI) and “Metric for Identifying a Blogger’s Influence” (MEIBIX) [62] are proposed. These metrics examine the temporal feature of a blogger’s activity and claim identification of the influential bloggers with respect to time. The metrics consider limited features like number of blogs initiated, the number of comments received and the number of links linking to these blogs. These metrics consider the time of the blog posts but lacks to consider the regularity and consistency of a blogger. The metrics extend to propose BI-index and BP-index metrics [66]. The BP-index evaluates a blogger’s and BI-index calculates their influence. The metrics lacks to introduce feature and are based on h-index, an index to rank authors and has certain shortcomings [67]. A main limitation is the lack of inclusion of all the comments and in-links and all the inlinks and comments except the h-core values become insignificant. Thus, two authors may have same h-index, but enjoy diverse number of inlinks and comments.
The MEIBI and MEIBIX metrics are extended by introducing features of FaceBookCount and uniqueness in a recent research work [68]. The reason of inclusion of new factor is to check whether the bloggers who have been identified influential in the blogosphere are also discussed in other social web channels or not. This assumption has been considered as an evaluation purpose. It takes into account the emotion of the blog content. It says that the model may be extended to add other social web platforms like Google+ score, Retweets, etc. A recent work proposes the blog ranking model by considering blog’s quality and temporal factors [69]. It considers a blogger’s impact as a measure and identifies the top blogs. Another research work presents [70] a model to identify top k-bloggers based on the idea of h-index, which is discussed earlier with its limitations. This work uses the same features as those of iFinder index [65] and computes the h-index of posts. The drawback of this work may be that as citation is the main feature to check the quality of a research publication and this taking this feature to calculate h-index is proper method and h-index is widely used metric to check the impact of a researcher but it may not be applied to features such as outlinks and post length. PostInfluence [71] is a metric to measure the quality of a blog post. Its influence score calculation is also based on iFinder, but it does not include the blogger’s self-comments. It takes introduces a model adopted from Pagerank [5] to find the top influential bloggers.

A new metric, BI-Impact, ranks web logs to find the top influential blogs within the blogosphere [72]. It takes into consideration varied features like the bloggers’ activity, post content, and the interaction of a post to calculate the influence of the blog. The features have also been assigned weights according to their significance. The structure of the blogosphere is explored to identify the influential blog user considering the six network centrality measures [73]. The scores of centrality measures are aggregated to find the influence score of a blogger. Longitudinal User Centred Influence model (LUCI) [74], considers network structure and explore bloggers’ interaction and classifies them into followers, extrovert and introvert leaders and neutrals. The optimal outcome of the model (90.3%) presents the significance of the features.

One of a current efforts [75] proposes approach using comments receive on post and consider iFinder [65] as baseline for results comparison. The authors discuss that the comment is an important feature and argue that iFinder model provides too much
Related Work

significance to inlinks. A recent model finds the top influencers in the social web using how much and how quickly their shared information is disseminated in the social forum [76]. It introduces InFlowMine algorithm to mine the information flow in the network. Another recent work ranks the top radically influential users of the social web [77]. It takes nine features in the form of metrics and then computes its weights and uses weighted PageRank algorithms to rank the top users, who are involved in malicious activities. This is a recent application of finding influential users in social web. A novel and recent model figures out how the influence of top influential users varies in a dynamic network.[78]. It uses the social network structure to find the top users who maintain their influence over a period of time.

The overview of existing works in this research domain is given in Table 3. Motivated by the limitations in the relevant literature, a new metric is proposed by introducing features, using modular approach and compare it with standard models and metric as given in chapter 3.

<table>
<thead>
<tr>
<th>Model/Metric</th>
<th>Classification</th>
<th>Features</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>iIndex [61]</td>
<td>Feature based Non-Temporal Model</td>
<td>number of inlinks, outlinks, length of post, number of comments</td>
<td>Unofficial Apple Weblog (TUAW)</td>
</tr>
<tr>
<td>WisClus clustering approach [63]</td>
<td>Feature based Non-Temporal Model</td>
<td>Same as iIndex, graph based approach</td>
<td>TUAW Digg &amp; BlogCatalog</td>
</tr>
<tr>
<td>SIIB (Semantically Identified Influential Bloggers) [64]</td>
<td>Feature based Non-Temporal Model</td>
<td>quantitative analysis of the contents, content semantics of blogposts</td>
<td>TUAW</td>
</tr>
<tr>
<td>MEIBI and MEIBIX [62]</td>
<td>Feature based Temporal Metric</td>
<td>Inlinks, comments and Temporal aspect</td>
<td>TUAW</td>
</tr>
<tr>
<td>iFinder [65]</td>
<td>Feature based Non-Temporal Model</td>
<td>Recognition, Activity Generation Novelty, Eloquence</td>
<td>TUAW</td>
</tr>
<tr>
<td>MEIBI-Index &amp; MEIBIX-Index [66]</td>
<td>Feature based Temporal Metric</td>
<td>Inlinks, Comments, and Temporal Aspect</td>
<td>Engadget</td>
</tr>
<tr>
<td>BP-index and BI-index [69]</td>
<td>Feature based Temporal Metric</td>
<td>Inlinks, Comments, and Temporal Aspect</td>
<td>Engadget</td>
</tr>
<tr>
<td>I-FBCount [68]</td>
<td>Feature based Non-Temporal Index</td>
<td>uniqueness and Facebook Count, sentimental factor, influence of comments</td>
<td>Engadget</td>
</tr>
<tr>
<td>h-index family Influence index</td>
<td>Post content, Inbound reference, Outbound links, Response, Related information</td>
<td>TUAW</td>
<td></td>
</tr>
</tbody>
</table>

“Social Web Mining based on Link and Content Analysis”
<table>
<thead>
<tr>
<th>[70]</th>
<th>number of reads.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PostInfluence (PInf) [71]</td>
<td>Feature centric Non-Temporal Model</td>
</tr>
<tr>
<td>BI-Impact [34]</td>
<td>Network Model</td>
</tr>
<tr>
<td>Six Centrality measures[35]</td>
<td>Network Model</td>
</tr>
<tr>
<td>LUCI [36]</td>
<td>Network Model</td>
</tr>
<tr>
<td>LCI [37]</td>
<td>Feature based Model</td>
</tr>
<tr>
<td>inFlowMine [80]</td>
<td>Network based model</td>
</tr>
<tr>
<td>RadicalUsers [81]</td>
<td>Network based PageRank</td>
</tr>
</tbody>
</table>

2.3 **Mixed-Opinion Classification of Web Forums**

Opinion mining includes subjectivity and sentiment classification. Sentiment classification categorized content into sentiment valence of either positive or negative. Neutral or no valence is also an option. Mixed- opinion content contains both positive and negative sentiments. Here we discuss the works where target was to find mixed-opinion and also review works for feature-centric sentiment analysis of online forums.

2.3.1 **Mixed-Opinion Content Classification**

The idea of analyzing positive and negative opinions is found in various studies. An early work analyzes the negative and positive sentiments in web [79]. Andre Bizau et al., [80] describes a natural language processing approach to identify diversity in opinions expressed in text. It focuses to capture both positive and negative reviews from the online forum. A study detects opinionated claims in online discussions using supervised learning methods using dataset of Wikipedia and LiveJournal [81]. It uses lexicon-based features for sentiment analysis and identifies users’ claims and arguments. Political text is analyzed to find the contrasting opinions of the users and to quantify their differences [82]. Another research work finds that contentions are important features of online political forums especially for those which have topics related to political, religious, and social issues. It also discovers that the agreement and contention indicator expressions at the post as well as thread level in the online forums [83].
Let us here describe related works that have diverse approach to find contrasting views and opinions which provide foundation to mixed opinion works. A natural language processing approach uses taxonomy of various types to identify contradictory claims within web content [84]. A corpus-based approach identifies co-occurrence patterns of words to measure the opinions of word [85]. Another similar approach finds the positive and negative words and then if their difference is near zero then it concludes that there is more contradictions in the text [86]. Using lexicon approach, pros and cons of political content has been used to identify the controversial political topics [87]. Another study focuses to find sentiment-based contradictions in web data by aggregating opinions [88]. An opinion diversification model detects the controversial topics after identifying positive, negative and neutral sentiments [89]. A statistical model finds sentiment diversity of comments using sentiment analysis [90]. The above-cited work present that either natural language approach using corpus and lexicon have mainly been used to identify opinionated content having both positive as well negative emotions.

### 2.3.2 Feature-Centric Classification of Online Forums

Various features-centric works use supervised learning techniques for opinion extraction from online forum content. We focus on use of thread features for various research works. The thread-structure has been exploited to categorize the user attitude towards other users. This supervised learning work helps to identify sentences which depict users’ attitudes, to finds dynamics of users’ interactions in threads and to identify formation of users’ groups and also their break-up in the social web forums [11]. Another classification classifies the online post content into the evaluative and non-evaluative sentences [15]. The content features classify the post content into agreement and disagreement ones [91]. The content and dialog features finds the existence of conversation in the forums [92]. Thread structure are found helpful to enhance the accuracy of posts retrieval. It proposed the expansion and count model by using the natural language features and weighting functions like distance and content similarity [93]. One of the recent works study analyses the subjectivity of online threads by proposing thread-structure features. They consider that the initial posts define the topic and subjectivity of the thread. They used two datasets, one from technical forum and
another from travelling forum defining it as subjective dataset [94]. The thread structure of the discussions in online forums has been exploited using the dataset of Reddit Community to evaluate the thread any creating hLDA clusters. It was found that the depth of early posts in a thread increases with the passage of time [12]. A recent study evaluates the subjective analysis of threads in online forum [94]. In this research, thread structure based features have been proposed to find the thread subjectivity orientation. The aforementioned discussion reveal that thread structure has been used for various research problems but no work that considers thread-specific features to find mixed-opinion content. Table 4 presents an overview of the existing works in the relevant literature.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Main Idea</th>
<th>Classification</th>
<th>Algo &amp; Features</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>[84]</td>
<td>Find contradictions in text</td>
<td>NLP</td>
<td>(Linguistic Features), Contradictions Corpus</td>
<td>Text Dataset RTE_Test</td>
</tr>
<tr>
<td>[85]</td>
<td>Find review as recommended or non-recommended.</td>
<td>NLP</td>
<td>Semantic orientation between opposite phrases</td>
<td>Eopinions review</td>
</tr>
<tr>
<td>[86]</td>
<td>Find Controversial topics by metrics of Provocativeness</td>
<td>Lexicon, NLP</td>
<td>Finding Opinion difference</td>
<td>Blog 06</td>
</tr>
<tr>
<td>[87]</td>
<td>Find Pro and Cons of Political Stance</td>
<td>NLP</td>
<td>Language Model</td>
<td>Political Forum</td>
</tr>
<tr>
<td>[88]</td>
<td>Discover contradictions on web.</td>
<td>NLP</td>
<td>Statistical measures, Track opinion over time</td>
<td>Slashdot and WebMD forums</td>
</tr>
<tr>
<td>[89]</td>
<td>Sentiment diversification of queries</td>
<td>Lexicon</td>
<td>Lexicon based sentiment identification</td>
<td>Google queries and retrieved documents data</td>
</tr>
<tr>
<td>[90]</td>
<td>Sentiment based Contradictions finding</td>
<td>Statistical Model</td>
<td>Lexical Features</td>
<td>DrugRatingz, YouTube, Slashdot Data</td>
</tr>
<tr>
<td>[14]</td>
<td>To find users attitudes is positive or negative discussions</td>
<td>Lexicon, NLP</td>
<td>Lexical patterns, POS patterns, Grammar patterns)</td>
<td>Usetnet online forum</td>
</tr>
<tr>
<td>[15]</td>
<td>To identify topical differences in posts of online discussions</td>
<td>Topic Model</td>
<td>LDA, clustering of documents</td>
<td>Reddit forum</td>
</tr>
<tr>
<td>[18]</td>
<td>Find evaluative and non-evaluative sentences</td>
<td>Lexicon</td>
<td>Lexical Features</td>
<td>Sino Web Data</td>
</tr>
<tr>
<td>[69]</td>
<td>Classify into Positive and negative</td>
<td>NLP</td>
<td>Linguistic Features</td>
<td>Web data</td>
</tr>
<tr>
<td>[70]</td>
<td>Positive and negative emotion identification</td>
<td>NLP</td>
<td>Linguistic Features</td>
<td>Web Forum</td>
</tr>
</tbody>
</table>
### Related Work

<table>
<thead>
<tr>
<th>Reference</th>
<th>Task Description</th>
<th>Techniques</th>
<th>Features</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>[71]</td>
<td>Find opinionated data</td>
<td>Lexicon, ML</td>
<td>Linguistic and Lexical Features</td>
<td>Wikipedia and LiveJournal</td>
</tr>
<tr>
<td>[72]</td>
<td>Find Contrasting opinions</td>
<td>ML</td>
<td>Linguistic Features</td>
<td>Political Forum Data</td>
</tr>
<tr>
<td>[73]</td>
<td>To Discover Agreement or Disagreement</td>
<td>Lexicon, ML</td>
<td>Lexical Features</td>
<td>Web Forum Post and Discussion</td>
</tr>
</tbody>
</table>
Chapter 3: Subjectivity Analysis using Non-Lexical Features
Chapter 3

Subjectivity Analysis using Non-Lexical Features

3.1 Introduction

The social web consists of a large number of online forums that contain user generated content in various domains such as health, travel, electronic products, sports and social issues. The social web users discuss topics related to practical life. As these forums contain personal experiences of a large number of people and people have facility of direct discussions, therefore, online forums are becoming popular for discussion real life issues.

The subjectivity analysis of web forum is the second research problem addressed. The posts and threads are categorized into either subjective or objective/non-subjective. This challenge has been taken as a binary classification model. The social web users discuss subjective issues in subjective discussions/threads, such as seeking personal views, sentiments, feedbacks and personal experiences, while non-subjective threads discuss factual and informative information. The table 5 illustrates the difference between subjective and non-subjective sentences:

<table>
<thead>
<tr>
<th>Subjective</th>
<th>Non-Subjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the quality of the new Smartphone camera?</td>
<td>What is the price of new Smartphone?</td>
</tr>
<tr>
<td>How safe is it to live in Karachi</td>
<td>What is the distance from Karachi to Lahore?</td>
</tr>
<tr>
<td>Hotel rates in living in New York</td>
<td>Hungary Brings in Tough Migration Laws</td>
</tr>
<tr>
<td>What does the future hold for our sprawling cities?</td>
<td>Which is the biggest City in Pakistan?</td>
</tr>
</tbody>
</table>

3.2 Why Subjectivity Analysis of Online Forums

The subjectivity analysis of threads on online forums helps to improve the forum search for a user who is interested to find single correct factual information and for another user who is keen to find various reviews from multiple users about a product. The knowledge
that a thread contains subjective or objective content helps to improve search in a forum. Due to the increased importance of online reviews, a lot of extraneous, off-topic, inflammatory and spam messages and reviews are incorporated in discussion threads. The content of such spam and unwanted messages is subjective in nature and is potentially helpful to analyze discussions for subjectivity. Web reviews are helpful and influential to web readers in establishing their points of views and sentiments on a certain topic.

A number of research works discuss the applications of subjectivity analysis. For instance, online reviews are analysed to rank products [95-97]. Subjectivity analysis of political forum is helpful to study public opinion polls [98], predict election results [25, 99], in medical domain [100, 101]. Apart of the social web sources, subjectivity and sentiment detection is detected other content such as emails [102], novels and fairy tales [103]. A recent survey provides extensive discussion about use of subjectivity analysis in various domains [104].

3.3 Problem Formulation and Problem Statement

The problem is formulated and stated as under.

3.3.1 Problem Formulation

A user in a web forum initiates a discussion or thread by creating a new post. The initial post content sets the topic for comments and discussion. The context of the discussion within a thread can be informative or opinionative. Thread topics can be objective or subjective. An objective topic contains consists of factual information and figures, whereas we define a subjective topic as defined in [81] and the one that seeks the public the personal views and sentiments. We assume that a thread discussion contains views about a single topic and context drift is not covered in this research. It is notable that topic can drift in the threads and our assumption may not hold right always but such exceptional cased are not out of the scope of this research.

Formally, we define a forum post \( p \) is a arrangement of words in a Vocabulary set \( V \), a forum thread \( t \) is a sequence of posts i.e., \( t = \{ p_1, p_2, ..., p_L \} \), where \( p_i \) is the \( i \)\(^{th} \) post in the thread and forum \( f \) to be a collection of \( m \) threads \( f = \{ t_1, t_2, ..., t_m \} \), where \( t_j \) is a thread. A thread is initiated by a user \( u \in U \) and \( U = \{ u_1, u_2, ..., u_n \} \) where \( n \) is the
users’ count. A post having high subjective content is regarded as subjective post, denoted as \( p_s \) otherwise it is non-subjective post, denoted as \( p_{ns} \). Similarly, a thread containing set of subjective discussion is taken as subjective threads, denoted as \( t_s \), otherwise, it is considered as non-subjective, denoted as \( t_{ns} \).

### 3.3.2 Problem Statement
Given an online forum \( f \) and the set of thread \( t \), our aim is to classify a post \( p \), into subjective \( p_s \) or non-subjective \( p_{ns} \) classes and a thread \( t \) into subjective \( t_s \), or non-subjective \( t_{ns} \).

### 3.4 Features Engineering
We propose dialog features and validate the importance of dialog features over conventional sentiment features for classification of posts in the social web forums. The research explores the effect of using lexical as well as non-lexical features for threads’ classification in either subjective or non-subjective categories with the help of standard classification algorithms. We describe the various types of features which have been used for subjective analysis. In this section, we propose both baseline and proposed post features. Here, we share in Table 6 the list of symbols used in this research module.

<table>
<thead>
<tr>
<th>Symbols ( T )</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>Set of Threads</td>
</tr>
<tr>
<td>( P )</td>
<td>Set of Posts</td>
</tr>
<tr>
<td>( U )</td>
<td>Set of Users</td>
</tr>
<tr>
<td>( t )</td>
<td>( t \in T )</td>
</tr>
<tr>
<td>( p )</td>
<td>( p \in P )</td>
</tr>
<tr>
<td>( u )</td>
<td>( u \in U )</td>
</tr>
<tr>
<td>( N_w^p )</td>
<td>Number of words ( w ) in post ( p )</td>
</tr>
<tr>
<td>( N_T )</td>
<td>Number of threads</td>
</tr>
<tr>
<td>( N_u )</td>
<td>Number of users</td>
</tr>
<tr>
<td>( N_{ur}^p )</td>
<td>Number of URLs in post ( p )</td>
</tr>
<tr>
<td>( N_{cc}^p )</td>
<td>Number of Capital Case words in post</td>
</tr>
<tr>
<td>( b_{un}^p )</td>
<td>Existence of Username mentioned in post</td>
</tr>
<tr>
<td>( b_{q}^p )</td>
<td>Existence of earlier thread posts quoted in post</td>
</tr>
<tr>
<td>( N_l^p )</td>
<td>Number of characters in post</td>
</tr>
</tbody>
</table>

### Table 6: List of Symbols used for Subjectivity Analysis

“Social Web Mining based on Link and Content Analysis”
3.4.1 Post Features

Here, post features are presented which are divided into lexicon-based sentiment and content-based dialog features.

3.4.1.1 Sentiment Features

These features consider post sentiment and are computed with the help of a lexicon. These features consider the emotion of a post. A post having opinion is likely to contain more subjective content as compared to an objective post that contains facts. The sentiment features are computed using various resources such as sentiment lexicon (for instance, SentiWordNet [30] and WordNet-Affect⁵) and sentiment analysis tool (e.g., SentiStrength [105, 106] and LIWC [107]). SentiWordNet has been used to compute sentiment features of posts content in this work. SentiWordNet is a widely used sentiment lexicon⁶ which computes the positive, negative and objectivity score of a content. SentiWordNet is compared with other lexicons and is found as the most suitable lexicon for sentiment classification [29]. It contains a large collection of terms as it is based on WordNet⁷, which is a world standard lexical dataset for the English dataset. The sentiment features have further been classified into Sentiment-Lexicon and Sentiment-WordFreq features. The Sentiment-Lexicon uses the SentiWordNet scores of the content while the Sentiment-WordFreq is proposed to compute the sentiment score considering the frequency of the sentiment words.

The negScore and posScore are the respective negative and positive sentiment scores of a post. These two features are the sentiment features. The features of numPosWords and numNegWords are the number of positive as well as negative word-count normalized by total words in post.

3.4.1.2 Dialog Features

Let us posit that an opinionative post is likely to have more conversation as compared to subjective post. An opinionative post possesses certain content features which depict the dialog among the thread users. It is anticipated that a thread of factual topic does not need

⁶ According to SentiWordNet Website statistics from google scholar, SentiWordNet has been used over 2000 research works. SentiWordNet. http://sentiwordnet.isti.cnr.it/
⁷ WordNet. https://wordnet.princeton.edu/
Subjectivity Analysis using Non-Lexical Features

a lot of content as no conversation is required, but subjective threads can have conversation and thus their content can have certain characteristics. We posit that conversation related features help to identify subjective thread. The first feature is the number of URLs (numURL) in the post content. A link is shared to redirect the conversation to another web page or another post within the forum which may be due to users posting links to other posts or web pages. We take that users share more links are shared to emphasize emphasizing their point of view opinion. The number of capital case words (numCapital) is also considered as harsh or emotional content. The content in capital case shows the shouting, or harsh words. A user mentions other users to seek the attention of a user of an earlier post. Similarly, a user copies content of an earlier post for a reply. To identify the dialog or conversation within the thread, user can mention the name of a user who has already commented on the topic. We consider it as feature; if a user mentions other user in his/her comment (boolUsername). Similarly, during conversation or replying exiting comment, exiting comment content may be copied as quoted content within a comment (boolQuotedText). We posit that factual information is short, while an opinionative content is likely to be lengthier. We consider post length as a feature. It is also argued that to add more argument, lengthy content may be posted by the user in the subjective thread (numCharacters). All the proposed post features are presented in table 7.

### 3.4.2 Thread Features

We propose features specific to thread-structure. Such features do not need a lexicon for computation. Let us again propose that a subjective thread has a higher chance of dialog than an objective thread. Thread structure features include the number of users having at least a post in thread, number of posts, thread life and thread length. To emphasize opinion, a user creates consecutive posts as we take it as a feature. To identify conversation, a user creates a post and another user creates a post in reply and the first user creates another post and comment-reply cycle is created. The count of such cycles is considered as a feature. The reason behind proposing thread features is that we consider that a subjective discussion takes more time, prompts more users to participate, urges the users to write more posts and have conversation among users. The table 8 shows the list
Subjectivity Analysis using Non-Lexical Features

of baseline thread features which have been proposed for the subjectivity analysis of online forums [94]. The list of proposed thread structure based features and proposed framework is given in Table 9 and Figure 2 respectively and algorithms are also presented. The time complexity of both the proposed algorithms is quadratic and which is consistent with the existing algorithms.

Figure 2: Layout of the Proposed Framework for Subjectivity Analysis

Table 7: List of Proposed Post Features for Subjectivity Analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiment Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posScore</td>
<td>$S_{p}^{p}$</td>
<td>Post’s Positive sentiment score (Lexicon Feature)</td>
</tr>
<tr>
<td>negScore</td>
<td>$S_{n}^{p}$</td>
<td>Post’s Negative sentiment score (Lexicon Feature)</td>
</tr>
<tr>
<td>numPosWords</td>
<td>$S_{pw}^{p}$</td>
<td>Number of Positive words over number of Post words (WordFreq Feature).</td>
</tr>
<tr>
<td>numNegWords</td>
<td>$S_{nw}^{p}$</td>
<td>Number of Negative words over number of Post words (WordFreq Feature)</td>
</tr>
<tr>
<td><strong>Dialog Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>numURL</td>
<td>$N_{ur}^{u}$</td>
<td>Number of URL present in a post</td>
</tr>
<tr>
<td>numCapital</td>
<td>$N_{c}^{p}$</td>
<td>Number of Capital Case words present</td>
</tr>
<tr>
<td>boolUsername</td>
<td>$b_{un}^{p}$</td>
<td>Existence of Username mentioned in a post</td>
</tr>
<tr>
<td>boolQuotedText</td>
<td>$b_{q}^{p}$</td>
<td>Existence of earlier thread posts quoted in a post</td>
</tr>
<tr>
<td>numChar</td>
<td>$N_{l}^{p}$</td>
<td>Number of Post characters</td>
</tr>
</tbody>
</table>
### Table 8: List of Baseline Thread Features for Subjectivity Analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>InitPostLength</td>
<td>Number of Words in First Post</td>
</tr>
<tr>
<td>ThreadLength</td>
<td>Number of Thread Words</td>
</tr>
<tr>
<td>AvgPostAuthor</td>
<td>Avg posts by a user</td>
</tr>
<tr>
<td>AvgLengthPost</td>
<td>Avg length of Thread posts (in words)</td>
</tr>
<tr>
<td>numUser</td>
<td>Number of users in the discussion</td>
</tr>
</tbody>
</table>

### Table 9: List of Proposed Thread Features for Subjectivity Analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>numConsPosts</td>
<td>$N_{cp}^t$</td>
<td>Number of Consecutive Posts by a user</td>
</tr>
<tr>
<td>NumCycleUsers</td>
<td>$N_{cu}^t$</td>
<td>Number of Users’ Dialog Cycles in a thread</td>
</tr>
<tr>
<td>numUsers</td>
<td>$N_u^t$</td>
<td>Number of users posting in a thread</td>
</tr>
<tr>
<td>numDays</td>
<td>$N_d^t$</td>
<td>The period (in days) from first to last post</td>
</tr>
<tr>
<td>numCharacters</td>
<td>$N_c^t$</td>
<td>Number of characters of all the posts in the thread</td>
</tr>
<tr>
<td>numPosts</td>
<td>$N_p^t$</td>
<td>Number of Posts in the thread</td>
</tr>
</tbody>
</table>

#### Non-Lexical Discrete Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>boolConsPosts</td>
<td>$b_{cp}^t$</td>
<td>Existence of Consecutive Posts by the same user</td>
</tr>
<tr>
<td>boolCycleUsers</td>
<td>$b_{cu}^t$</td>
<td>Existence of Dialog Cycles of users in the thread</td>
</tr>
<tr>
<td>boolUsers</td>
<td>$b_u^t$</td>
<td>Number of the thread users $&gt; \text{Average users in all the threads}$</td>
</tr>
<tr>
<td>boolDuration</td>
<td>$b_d^t$</td>
<td>Number of Thread days $&gt; \text{Average days of all the threads}$</td>
</tr>
<tr>
<td>boolLengthyThread</td>
<td>$b_{lt}^t$</td>
<td>Length of thread $&gt; \text{Average Length of all the threads}$</td>
</tr>
<tr>
<td>boolMorePosts</td>
<td>$b_p^t$</td>
<td>Number of Thread posts $&gt; \text{Average number of posts in all the threads}$</td>
</tr>
</tbody>
</table>

### Mathematical Formulas

$$b_u^t = \begin{cases} 1, & \text{if } N_u^t > \frac{N_T}{N_U} \\ 0, & \text{otherwise} \end{cases}$$

$$b_d^t = \begin{cases} 1, & \text{if } N_d^t > \frac{N_T}{N_D} \\ 0, & \text{otherwise} \end{cases}$$

$$b_{lt}^t = \begin{cases} 1, & \text{if } N_{lt}^t > \frac{N_T}{N_L} \\ 0, & \text{otherwise} \end{cases}$$
ALGORITHM 1: Subjectivity Analysis of Posts in Online Forum

**Input:** Data of Posts in Online Forum

**Output:** Post classified as subjective or non-Subjective

1. Initialize $N_w^P, N_{pw}^P, N_{nw}^P, N_{ur}^P, N_c^P, N_l^P$
2. FOR each $t \in T$
3. FOR each $p \in P$
4. $N_w^p = \text{CountWords}(p)$
5. $N_{pw}^p = \text{CountPositiveSWNWords}(p)$
6. $N_{nw}^p = \text{CountNegativeSWNWords}(p)$
7. \(\Rightarrow\) Computation of Sentiment Feature Set ($F_S^P$)
8. $S_p^p = \text{CalculatePositiveSWNScore}(p)$
9. $S_n^p = \text{CalculateNegativeSWNScore}(p)$
10. $S_{pw}^p = N_{pw}^p / N_w^p$
11. $S_{nw}^p = N_{nw}^p / N_w^p$
12. $F_S^p = [S_p^p; S_n^p, S_{pw}^p, S_{nw}^p]$
13. \(\Rightarrow\) Computation of Post Dialog Feature Set ($F_D^P$)
14. IF $p$ contains URL THEN
15. $N_{ur}^p = N_{ur}^p + 1$
16. END IF
17. IF $p$ contains Capital Word THEN
18. $N_c^p = N_c^p + 1$
19. END IF
20. IF $p$ contains Username of Earlier posts in the thread $t$ THEN
21. $b_{un}^p = 1$
22. ELSE
23. $b_{un}^p = 0$
24. END IF
25. IF $p$ contains Quoted text from Earlier posts in the thread $t$ THEN
26. $b_q^p = 1$
27. ELSE
28. $b_q^p = 0$
29. END IF
30. $N_l^p = \text{CountWords}(p)$
31. $F_D^p = [N_{ur}^p, N_c^p, b_{un}^p, b_q^p, N_l^p]$
32. END FOR
33. END FOR
34. $Class = \text{Classifier}(F_S^P, F_D^P)$
35. IF $Class = 1$ THEN $p' = p_s$
36. Else $p' = p_{ns}$
37. END IF
38. STOP \(\Rightarrow\) END of Algorithm
ALGORITHM 2: Subjectivity Analysis of Threads in Online Forum

Input: Data of Threads in Online Forum
Output: Thread classified as subjective or non-Subjective

1. Initialize $N_u^t, N_p^t, N_{cp}^t, N_{cu}^t, N_u^t, N_d^t, N_c^t$.
2. FOR each $t \in T$
3. ▶ Computation of Thread Discrete Feature Set ($F_d^t$)
4. $N_u^t = N_u^t + 1$
5. $N_p^t = N_p^t + 1$
6. IF $u_{i-1}^p = u_i^p$
7. $N_{cp}^t = N_{cp}^t + 1$
8. END IF
9. IF $u_{i-1}^p = u_{i+1}^p$
10. $N_{cu}^t = N_{cu}^t + 1$
11. END IF
12. $N_u^t = \text{CountWords}(t)$
13. $N_p^t = \text{CountDas}(t)$
14. $F_d^t = [F_d^p; N_p^t, N_{cp}^t, N_{cu}^t, N_u^t, N_d^t, N_c^t]$
15. ▶ Computation of Thread Boolean Feature Set ($F_b^t$)
16. IF $N_{cp}^t > 0$ THEN $b_{cp}^t = 1$
17. Else $b_{cp}^t = 0$
18. END IF
19. IF $N_{cu}^t > 0$ THEN $b_{cu}^t = 1$
20. Else $b_{cu}^t = 0$
21. END IF
22. IF $N_u^t > \frac{N_U}{N_T}$ THEN $b_u^t = 1$
23. Else $b_u^t = 0$
24. END IF
25. IF $N_d^t > \frac{N_D}{N_T}$ THEN $b_d^t = 1$
26. Else $b_d^t = 0$
27. END IF
28. IF $N_c^t > \frac{N_C}{N_T}$ THEN $b_c^t = 1$
29. Else $b_c^t = 0$
30. END IF
31. IF $N_p^t > \frac{N_P}{N_T}$ THEN $b_p^t = 1$
32. Else $b_p^t = 0$
33. END IF
34. $F_b^t = [b_{cp}^t, b_{cu}^t, b_u^t, b_c^t, b_p^t, b_d^t]$
35. END FOR
36. $Class = \text{Classifier}(F_d^t, F_b^t)$
37. IF $Class = 1$ THEN $t' = t_s$
38. Else $t' = t_{ns}$
39. END IF
40. STOP ▶ END of Algorithm

3.5 Experimental Setup

In this section, classification algorithms applied, the dataset used and performance evaluation measures used are briefly given.
Subjectivity Analysis using Non-Lexical Features

3.5.1 Classification Algorithms

The Oracle Data Miner [108, 109] is an advanced tool for data mining. It provides various data analysis and mining algorithms for classification, feature selection, prediction etc. ODM is used for various data mining tasks for various applications such as predictive analysis [110], disease classification [111], educational data mining [112]. Four varied classification algorithms, i.e., Naïve Bayes, Decision Tree, Support Vector Machine and Logistic Regression are applied using ODM. It uses Maximum Description Length (MDL) algorithm for computing whether the attribute role is positive or not. All the proposed features show positive results. The ODM uses k-fold cross validation techniques and we have used 10-cross validation. Experiments apply the ODM default settings. The ODM is freely available for research purposes\(^8\). Features are normalized between 0 and 1. Supervised learning techniques is widely used approach which produced promising results in subjectivity and opinion mining [19, 37, 55, 94, 113].

3.5.2 Data

The choice of forum dataset is significant as it should have topics including both factual and subjective and content should be generated by diverse from worldwide. We have used data of BBC Forum dataset [114], a public discussion forum, provides the target values of subjective and objective. The users discuss various topics in a thread from general news, social issues to political and religious views etc.

3.4.2.1 Data Analysis of BBC Dataset

A relationship exists between the properties of a dataset, which are represented in the form of variables. A relationship can wither be linear or non-linear. If a value of a quantity alters as a power of second quantity value, then such a relationship is expressed using power law. The power law\(^9\) is used to study the probability distributions. The distributions of a large data of physical, biological, and human related fields of life follow the power law distributions. For instance, the frequencies of words in most languages, the ranges of earth quakes etc., observe power law form. For the analysis of the number of


users and their participation in the social web forum, power law analysis has been used as well. The number of posts or messages in threads varies following power law as well. It has been followed using the various research works [115]. It has been observed that the empirical power-law distributions hold for a limited range and a number of such values observe the law but others fir the power law in the tail.

In order to represent the uncertainty in the experimental values, a deviation term $\epsilon$ is added in the power law $O(x^k)$ or simply $y = ax^k + \epsilon$ for observing the value of a deviation from the main function of the power law. The power law distribution is given as follows:

$$p(x) = cx^{-x} \quad \text{For } x > x_{\text{min}} \quad (1)$$

If the observed value of $\propto$ is found to be higher than unity, then the tail spread across an infinite region. It is required that the minimum value ($x_{\text{min}}$) should be as less as possible. The constant value of co-efficient $C$ is the scaling factor that helps to make sure the total are is 1, which is the requirement of the probability distribution.

To study the overall data distribution in the data of BBC over all the five years of the dataset, the statistical analysis is carried out on user’s posts quantities. The objective of the analysis is to observe the relationship between the user behavior and the distribution of posts over the years. Different users have different behavior in threads of online forum threads. An active user may share more comments as compared to other inactive users and this behavior is usually common in all the social web forums.

Let us consider a user activity, referred as $a_i$, defined as the count of the posts of a user $i$ and this is represented as $a$. The maximum number of the posts by a user in BBC dataset is, i.e., $a_{\text{max}} = 18,289$. It represents that the single user has shared more than eighteen thousand messages. On the other hand, the average posts per user or the average activity is $<a_{\text{avg}}>$ = 136, and the median value of user activity is $a_{\text{med}}$ = 3, which is very less and suggest that the majority of the users remain very less active in the forum. The observed statistics are given in Table 10 which shows the comparison of the BBC dataset with the datasets of Ubuntu and NYC which are used by our baseline research work [94, 113]. The similar trend is observed for the other two recent datasets as well.
Table 10: A Comparision of Dataset Characteristics

<table>
<thead>
<tr>
<th>Dataset Features</th>
<th>BBC</th>
<th>Ubuntu</th>
<th>NYC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users Count</td>
<td>18017</td>
<td>1821</td>
<td>1151</td>
</tr>
<tr>
<td>Maximum number of messages by a single user $a_{max}$</td>
<td>18289</td>
<td>22</td>
<td>366</td>
</tr>
<tr>
<td>Average posts in threads $a_{avg}$</td>
<td>136</td>
<td>1.73</td>
<td>5.64</td>
</tr>
<tr>
<td>Median $a_{med}$</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Value of $\beta$</td>
<td>2.04</td>
<td>0.53</td>
<td>0.99</td>
</tr>
<tr>
<td>Value of Best Fit $R^2$</td>
<td>0.950</td>
<td>0.899</td>
<td>0.933</td>
</tr>
</tbody>
</table>

The power law analysis of the BBC dataset shows that the dataset follows the normal distributions as majority of the datasets. The value of $R^2$, which is known as the “co-efficient of Regression”, represents the accuracy of the curve with respect to the data. It ranges from 0 to 1. The higher values depict the accurate result or the fitness of the curve. Its value for BBC dataset is 0.95 which reveal the very high level of accuracy or “best fit” with respect to its data. Similarly, the value of $\beta$ shows high number of active users of the data. Its value is relatively higher for our dataset which reveal that the users activity remains similar in recent and BBC dataset as on average more number of users are active in the forum as compared to other datasets used for sentiment analysis and mining.

![Power Law Analysis of BBC Dataset](image)

**Figure 3:** Power Law Analysis of BBC dataset using Log Scale Chart
### 3.5.3 Performance Evaluation Measures

The standard performance evaluation measures of Precision, Accuracy, Recall and F-Measure have calculated using the following equations:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (3)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (4)
\]

\[
F - \text{Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)
\]

Where \( TN, TP, FN, FP \) represents the respective values of “True Negative, True Positive, False Negative and False Positive”.

### 3.6 Results and Discussion

Here, we present post and thread results using varied machine learning techniques and compare the feature sets various performance evaluation measures.

#### 3.6.1 Post Results

The post results of applied classification algorithms are presented in Tables 11. The result analysis reveals that dialog features outperform sentiment features. The sentiment lexicon features have shown better results as compared to the sentiment Word-Frequency features. The combination of Sentiment and Dialog features shows promising results.

Comparing the classification algorithms used, Logistic regression show optimal results for sentiment features, Support vector machines and Decision Tree show better results for dialog features. Decision tree, SVM and logistic regression all present optimal results using both sentiment and dialog features. We consider the optimal result achieved using any classification method and the results are given in Figure 4 which shows the significance of dialog features.
Table 11: Feature Set-wise Post Results for subjectivity analysis

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Algo</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiment</strong></td>
<td>NB</td>
<td>0.537</td>
<td>0.524</td>
<td>0.635</td>
<td>0.574</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.53</td>
<td>0.523</td>
<td>0.707</td>
<td>0.602</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.518</td>
<td>0.522</td>
<td>0.474</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.62</td>
<td>0.6</td>
<td>0.729</td>
<td>0.658</td>
</tr>
<tr>
<td><strong>Dialog</strong></td>
<td>NB</td>
<td>0.633</td>
<td>0.598</td>
<td>0.821</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.641</td>
<td>0.604</td>
<td>0.831</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.631</td>
<td>0.585</td>
<td>0.914</td>
<td>0.713</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.632</td>
<td>0.59</td>
<td>0.88</td>
<td>0.706</td>
</tr>
<tr>
<td><strong>Sentiment + Dialog</strong></td>
<td>NB</td>
<td>0.644</td>
<td>0.604</td>
<td>0.845</td>
<td>0.705</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.639</td>
<td>0.596</td>
<td>0.869</td>
<td>0.707</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.64</td>
<td>0.598</td>
<td>0.86</td>
<td>0.706</td>
</tr>
</tbody>
</table>

Figure 4: Post Classification Results for each Feature set

3.6.2 Thread Results

The thread results reveal that thread-structure helps to classify the subjective threads and non-subjective threads as per analysis from Table 12. Boolean features and numeric features depict the existence and count of the target characteristics of the thread. The results show that both categories using combined approach provide optimal results.
Comparing the classification algorithms, for numeric features, Decision tree and Logistic regression have shown higher results while for Boolean features, Logistic regression and Naïve Bayes show better results. Combining both categorizes Logistic regression shows better results as well. It is notable from Figure 5 that for all the proposed features present comparable results and no striking difference are found. Boolean features are derived from numeric features, but it is interesting that these show similar results.

### Table 12: Thread Classification Results for Subjectivity Analysis

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Algo</th>
<th>Accu</th>
<th>Prec</th>
<th>Rec</th>
<th>F-Meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>NB</td>
<td>0.715</td>
<td>0.688</td>
<td>0.723</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.682</td>
<td>0.627</td>
<td>0.701</td>
<td>0.705</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.656</td>
<td>0.615</td>
<td>0.702</td>
<td>0.712</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.762</td>
<td>0.698</td>
<td>0.801</td>
<td>0.799</td>
</tr>
<tr>
<td>Boolean</td>
<td>NB</td>
<td>0.705</td>
<td>0.633</td>
<td>0.811</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.714</td>
<td>0.615</td>
<td>0.800</td>
<td>0.714</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.642</td>
<td>0.621</td>
<td>0.872</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.726</td>
<td>0.689</td>
<td>0.846</td>
<td>0.778</td>
</tr>
<tr>
<td>All</td>
<td>NB</td>
<td>0.724</td>
<td>0.717</td>
<td>0.822</td>
<td>0.741</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.692</td>
<td>0.699</td>
<td>0.803</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.662</td>
<td>0.723</td>
<td>0.798</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.771</td>
<td>0.786</td>
<td>0.833</td>
<td>0.825</td>
</tr>
</tbody>
</table>

Figure 5: Thread Results for Subjectivity Analysis

### 3.7 Most Useful Features

We here find the importance of each feature, according to its role in the classification model. For this purpose, we applied the Wald chi-square statistic, $W_f$, which is computed for a feature $j$, $f_j$, as follows:
Subjectivity Analysis using Non-Lexical Features

\[ W_j = \frac{f_j}{SE_{bj}} \]  

(6)

Where \( W \) is the Wald’s statistic having normal distribution, \( f \) is the feature (coefficient) and \( SE \) is the standard error. The value of Wald’s statistic is squared to yield a Wald statistic with a chi-square distribution.

\[ SE_j = sqrt(diag(H^{-1})_j) \]  

(7)

Where

\[ H = [X^TWX] \]  

(8)

The ranking of each feature within the feature set for the post and thread is shown in tables 13 and 14 respectively.

For sentiment, the negative score is the top ranked features, which suggests that negative content plays an important role than positive sentiment content for subjectivity analysis. Our finding is similar as [114] which suggests that negative emotions causes additional content generation in online forums. It is notable that for SentiWordNet based features, the score features are more helpful as compared to the word features, which is understandable as the former shows the strength of the sentiment valence while the former counts the number of sentimental words. In dialog features, capital content is ranked high as depicts strong negative emotion usually [18]. The features of the user’s mentioning (boolUsername and boolQuotedText) have high ranks as both depicts the direct conversation among the users which is more common in subjective as compared to objective content. The post length is important already found in opinion mining research [18]. The feature to detect out-link (boolURL) does not enjoy high rank as a user may share a link in an informative post or link to another content which may be factual for more description of a factual information.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Dialog</th>
</tr>
</thead>
<tbody>
<tr>
<td>negScore</td>
<td>boolCapital</td>
</tr>
<tr>
<td>posScore</td>
<td>boolUsername</td>
</tr>
<tr>
<td>numNegWords</td>
<td>boolQuotedText</td>
</tr>
<tr>
<td>numPosWords</td>
<td>numChar</td>
</tr>
<tr>
<td></td>
<td>boolURL</td>
</tr>
</tbody>
</table>
From the table 14, it is notable that the number of posts is ranked high it shows that subjective threads are higher as compared to objective threads. The subjective discussions are lengthier as compared to informative discussion. The feature to capture users’ conversation (numCycleUsers) is ranked high that suggests that conversation is an important point. The thread length is significant feature which is a consistent observation as discussed in post features as well. The consecutive posts feature is low in rank as a user may post in answering an objective questions and sharing more information in response. The number of users and thread duration are ranked low as the threads users have the possibility to post any time irrespective of the nature of subjective or objective content. , thus the duration may lengthy or the emotional discussion may be done in a short span of time.

<table>
<thead>
<tr>
<th>Discrete</th>
<th>Boolean</th>
</tr>
</thead>
<tbody>
<tr>
<td>numPosts</td>
<td>boolMorePosts</td>
</tr>
<tr>
<td>numCycleUsers</td>
<td>boolCycleUsers</td>
</tr>
<tr>
<td>numChars</td>
<td>boolConsPosts</td>
</tr>
<tr>
<td>numConsPosts</td>
<td>boolLentghyThread</td>
</tr>
<tr>
<td>numUsers</td>
<td>boolUsers</td>
</tr>
<tr>
<td>numDays</td>
<td>boolDuration</td>
</tr>
</tbody>
</table>

### 3.8 Chapter Summary

In this chapter, subjective analysis of online forum has been carried out. First we present the lexicon-based and dialog based content feature sets for posts in the online forums. We show that the proposed dialog based features which need no lexicon outperform the conventional lexical features. Then for subjectivity analysis of threads in the social web forum, the thread features have been introduced. The thread features have been categorized into discrete and Boolean features. It is noteworthy that the proposed non-lexical features are very helpful as evident from the use of Boolean features as well. Boolean features measure the existence of a certain factor within the whole thread. We took novel approach that Boolean features also are based on the average value of the whole dataset as evident from the algorithms. The results confirm that the non-lexical thread specific feature set outperforms lexical feature set to separate subjective content from non-subjective content.
Chapter 4: Identification of Influential Bloggers
Chapter 4

Identification of Influential Bloggers

4.1 Introduction

This chapter covers the research work about finding top influential bloggers in the Blogosphere. We initially provide problem formulation and statement. The list of proposed features is discussed to propose the Metric for Identification of Influential Bloggers (MIIB). The results discussion includes feature-wise and module-wise analysis and a comparison of the proposed MIIB versus the existing model and metrics. The use of performance evaluation measures helps to compare the proposed metric.

4.2 Why Finding Influential Bloggers

Finding the top influential bloggers is a significant research problem. The goal in technical blogs is to look for quality content generated by specialists and main focus in marketing blogs is to discover trustworthy customers and find their feedback and experiences. The companies look for effective users who may become indirect representatives for marketing and product uplift. Such bloggers are market movers and trend setters as may help other users to analyse trends in marketing, product, technology, fashion, etc. [116, 117]. A study presents that the advertisements stimulate the blogging activity [118], which shows that marketing is related to online blogging activities. Blogs have already reshaped the face of journalism and are considered as influential actors for general social issues, political campaigns and policy as well as agenda settings [119-121].

A study focuses on the collective actions of community bloggers and presents deeper insights about influential community members’ roles in socio-political movements [122, 123]. The social web has also touched the social and personal lives of million users and finding the influential bloggers is an important tool. A recent study concludes that the influential bloggers play their role in health awareness campaigns such as anti-smoking [124], healthy properties[125] and online activities related to health and social issues [126].
4.3 Problem Formulation and Problem Statement
Here, we provide the problem formulation and problem statement.

4.3.1 Problem Formulation
A blogger $b \in B$ initiates a blog post $p$ and others post their views in the form of comments. A blog that attracts bloggers is an influential blog. A post that motivates bloggers to post a comment or generate an in-link is referred as an influential post. So, a blogger who initiates such posts is regarded as an influential blogger. We aim to identify the set $I$ containing top $k$ influential blog users based on their influence scores, $S_{infl}$, computed using certain features. The weight of a feature depicts its significance.

4.3.2 Problem Statement
Given a set $B$ of $N$ bloggers, $\{b_1, b_2, ..., b_N\}$, the problem of identifying the influential bloggers is formally, finding a set of bloggers $I$ ordered according to their influence scores, $S_{infl}$, such that $I \subseteq B$ and $K \leq N$, i.e., $S_{infl}(b_{j1}) \geq S_{infl}(b_{j2}) \geq ... \geq S_{infl}(b_{jk})$. The set $I$ contains the $k$ most influential bloggers.

4.4 MIIB: The Proposed Metric
This section presents a list of factors, the modules and the metric. The symbols used are displayed in Table 15.

4.4.1 Features Measuring the Blogger’s Influence
The list of all the features is presented in table 15 and each is elaborated as below:

*Activity (f1):* The capability of a blogger to create new posts is a significance factor. It is represented by $N_p^b$. This feature is considered widely in the relevant literature [61, 65, 66, 68, 71, 72].

*Activeness (f2):* A blogger ought to be active to be effective. A blogger may create too many posts during a short period but becomes inactive for a large period. It is discussed that an active blogger creates impact on others [72] so it is introduced as a factor. Activeness is the count of days a blogger remains active. It is represented as $N_d^b$. 
### Table 15: List of Symbols used for the Proposed Metric

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>A set of Bloggers</td>
</tr>
<tr>
<td>$P$</td>
<td>A set of Blog Posts</td>
</tr>
<tr>
<td>$S$</td>
<td>A set of blog Sites</td>
</tr>
<tr>
<td>$b$</td>
<td>$b \in B$</td>
</tr>
<tr>
<td>$p$</td>
<td>$p \in P$</td>
</tr>
<tr>
<td>$s$</td>
<td>$s \in S$</td>
</tr>
<tr>
<td>$N^b_p$</td>
<td>No of posts of a blogger $b$</td>
</tr>
<tr>
<td>$N^b_d$</td>
<td>No of days of posting of a blogger $b$</td>
</tr>
<tr>
<td>$S^b_r$</td>
<td>Score of a blogger’s regular posting behavior</td>
</tr>
<tr>
<td>$N^b_l$</td>
<td>PostLength of a blogger’s posts</td>
</tr>
<tr>
<td>$S^b_a$</td>
<td>Score of Avg-length of posts of a blogger $b$</td>
</tr>
<tr>
<td>$N^b_c$</td>
<td>Comments on a blogger’s posts</td>
</tr>
<tr>
<td>$N^b_i$</td>
<td>Inlinks received on a blogger’s posts</td>
</tr>
<tr>
<td>$N^b_o$</td>
<td>Outlinks in blog posts posted by a blogger</td>
</tr>
<tr>
<td>$N^b_h$</td>
<td>Number of Bloggers who post in a blog site $s$</td>
</tr>
<tr>
<td>$N^s_p$</td>
<td>Posts in a blog site $s$</td>
</tr>
<tr>
<td>$N^s_i$</td>
<td>In-links received by posts in a blog site $s$</td>
</tr>
<tr>
<td>$N^s_c$</td>
<td>Number of comments received by posts in a blog site $s$</td>
</tr>
<tr>
<td>$S^p_{prod}$</td>
<td>Productivity score using productivity features</td>
</tr>
<tr>
<td>$S^p_{popu}$</td>
<td>Popularity score using popularity features</td>
</tr>
<tr>
<td>$S^b_{Rank}$</td>
<td>Score of blogsite rank features</td>
</tr>
<tr>
<td>$S^s_{Rank}$</td>
<td>Score of a blogsite using blogsite rank features</td>
</tr>
<tr>
<td>$S^b_{Infl}$</td>
<td>A blogger’s Influence Score</td>
</tr>
</tbody>
</table>

**Consistency (f3):** Consistency measures that a blogger’s regular posting ability. The relevant literature [72] argue that a blogger be consistent to have influence as influence vanish with time. The existing work [17, 62, 66, 68] shares that time is an important factor. It is based on period between the consecutive posts of a blogger. It is represented as $S^b_r$, and is computed by dividing posts count by the posting period (measured in months) using equation 1:

$$\text{Consistency} = \frac{N^b_p}{(\max(\text{postdate}) - \min(\text{postdate})/30)} \quad (9)$$

**Recognition (f4):** Existing methods considers the comments received on a blogger’s posts as an important factor as it shows their recognition in the blogging community. It is denoted as $N^b_c$. 

“Social Web Mining based on Link and Content Analysis”
Identification of Influential Bloggers

**Authority** (f5): Ranking algorithms [5] which are based on graph-based social networks consider the in-links as a factor of authority. Relevant literature also consider it as the important feature [66]. It is denoted as $N^b_i$.

**Novelty** (f6): An out-link depicts the less novelty in a blog content. It is represented by $N^b_o$. As an inverse factor, it is aim to find such users having high post count, but less number of out-links. Considering less out-links results in bloggers having very less number of even no posts.

**BlogRank** (f7): Let us posit that bloggers to be influential if they post on top blog sites. We initially look for the top blog and later the blogger who uses top blog sites. It is represented as $S^s_{BRank}$. Using the blog site score, the top bloggers are identified.

**PostLength** (f8): The post length, in relevant literature, is considered as a measure to show the quality of content and is represented by $N^b_l$. We take number of characters in a post’s content.

**NormalizedPostLength** (f9): It is reasoned that a blogger may share too lengthy content and it will assign him higher score. To overcome, we introduce the normalized length also. This feature, represented as $S^b_{ab}$, is computed by dividing the count of a bloggers’ posts length by his/her number of posts.

The factors and their purposes are enlisted in Table 16.

<table>
<thead>
<tr>
<th>Feature no</th>
<th>Feature</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>Activity</td>
<td>A blogger’s post creating ability</td>
</tr>
<tr>
<td>f2</td>
<td>Activeness</td>
<td>A blogger’s ability to be active</td>
</tr>
<tr>
<td>f3</td>
<td>Consistency</td>
<td>A blogger’s consistent posting behavior</td>
</tr>
<tr>
<td>f4</td>
<td>Recognition</td>
<td>A blogger’s recognition in a blogging community</td>
</tr>
<tr>
<td>f5</td>
<td>Authority</td>
<td>A blogger’s recognition in a blogging community</td>
</tr>
<tr>
<td>f6</td>
<td>Novelty</td>
<td>Measures a blogger’s content novelty</td>
</tr>
<tr>
<td>f7</td>
<td>BlogRank</td>
<td>A blogsite’s significance in the blogosphere</td>
</tr>
<tr>
<td>f8</td>
<td>PostLength</td>
<td>Eloquence of a blogger’s content</td>
</tr>
<tr>
<td>f9</td>
<td>NormalizedPostLength</td>
<td>Average quality of a blogger’s content</td>
</tr>
</tbody>
</table>
The proposed framework is presented in Figure 6 as follows:

![Diagram of the Proposed Framework for MIIB]

**Figure 6: Layout of the Proposed Framework for MIIB**

### 4.4.2 The Modules of MIIB

All the existing models take linear approach, but we introduce a modular approach to depict the importance of modules. MIIB consists of popularity, productivity and BlogRank.

#### 4.4.2.1 Popularity Score

Popularity refers to a blogger’s impact on other users of the blogging community. It considers the comments and in-links received from other users in the community. It is debated that a comment in a blog post may not necessarily be in favor. It can be against the initial blog post which serves as the topic of the blog creator. An in-link depicts a direct impact whereas an outgoing link is considered reversely proportional to the content novelty. The number of outlinks are deducted from inlinks. The popularity score is computed using following equation 10:

Equation 10:
Identification of Influential Bloggers

$$S_{\text{popu}}^b = w_c N_c^b + (w_1 N_1^b - w_0 N_0^b)$$  \hspace{1cm} (10)

Where \(w_c, w_1\) and \(w_0\) denotes the weights of comments, in-links and out-links correspondingly.

### 4.4.2.2 Productivity Score

A blogger who initiates blog posts on consistent and regular basis is considered as productive. It takes into account a blogger’s features of activity, activeness and consistency. A blogger’s activity means blog creating capability which is the important feature \([62, 65, 68]\), whereas remaining features depend on it. iFinder considers only the post length as content eloquence measure. It is argued that a small number of too lengthy posts may assign unfair score thus the feature NormalizedPostLength introduces the concept of averaged length. The productivity score is calculated using the equation 11:

$$S_{\text{prod}}^b = w_p N_p^b + \left( w_d N_d^b + w_r S_r^b \right) + \left( w_1 N_1^b + w_a S_a^b \right)$$  \hspace{1cm} (11)

Where \(w_p, w_d\) and \(w_r\) denotes the weights of activity, the activeness and consistency correspondingly and \(w_1, w_a\) are PostLength weight and normalizedPostLength weight.

### 4.4.2.3 Blog Quality Score

Let us propose that blogsite is an important feature. MIIB includes blog site as an important factor and its score is calculated using equation 12:

$$S_{\text{BRank}}^s = (N_{bi}^s + N_{pi}^s + N_{Ii}^s + N_{Ci}^s)$$  \hspace{1cm} (12)

Where \(S_{\text{BRank}}^s\) represents the blog sites rank calculated using equation 4. The top bloggers having higher count of blog posts on the top weblog sites and its score is denotes as \(S_{\text{BRank}}^b\). It is notable that weights are not assigned to features in equation 4, as all the features depicts the overall importance of a blogsite and later the top bloggers are identified.

### 4.4.3 The Influence Score

The overall score of a blogger’s influence, \(S_{\text{infl}}^b\), uses all the three scores using weighted accumulative technique as given in equation 13:

$$S_{\text{infl}}^b = w_{\text{prod}} S_{\text{prod}}^b + w_{\text{popu}} S_{\text{popu}}^b + w_{\text{bank}} S_{\text{BRank}}^b$$  \hspace{1cm} (13)
Where $w_{\text{prod}}$, $w_{\text{popu}}$ and $w_{\text{bank}}$ refer to the respective weights for productivity, popularity and BlogRank modules.

### 4.4.4 Use and Effect of Weights

The feature weights regulate the role of features in the computation of the influence score. The weights adjustment approach is based on iFinder method. In popularity module, using the equation, we fix two weights and vary third from 0 to 1 and observe how the ranking changes. Fixing weights of inlinks $w_i$ and outlinks $w_o$, the comments weight stabilizes $w_c \geq 0.6$. While varying in-links weight and out-links weights the model stabilizes for inlinks weight $w_i \geq 0.9$ and out-links weights $w_o \geq 0.2$. The normalized weights for inlinks, comments and outlinks are 0.6, 0.3 and 0.1 respectively. It suggests the significance of inlinks which are also given higher weights in the existing works [62, 66, 72]. The statistics presented in Table 17 approve that the inlinks are more important than as compared to comments.

A similar approach finds the weights applied in other two modules. In productivity module, $w_p$, activity weight is 0.8, the $w_d$ and $w_r$ are found as 0.3 while the weights related to post length are 0.2. The normalized weights for activity is 0.4, consistency and activeness is 0.2 and potlength and normalizedPostLength are 0.1. It validates our assumption that activity is the important characteristics to measure a blogger’s productivity while other features depend on activity. For a blogger’s influence score, the modules weights for productivity $w_{\text{prod}}$, popularity $w_{\text{popu}}$ are found 0.4 respectively and value of $w_{\text{bank}}$ is 0.2.

## 4.5 Experimental Setup

This section presents an introduction to the dataset and performance evaluation measures used. The algorithm is given as follows:

### 4.5.1 TUAW Dataset

Apple commenced its world famous blog, The Unofficial Apple Weblog (TUAW), for publishing information covering topics related to its products and for providing guidance and help to its customers and for marketing purposes as well. TUAW used to offer
opportunity for bloggers to post comments, share feedbacks and experiences regarding the topics under discussion in the posts. The blog is currently closed\(^\text{10}\). The dataset has been used widely in the relevant literature [61-63, 65, 70] and is available free of cost\(^\text{11}\). The dataset comprises of five years data. The statistics explaining the characteristics of the TUAW dataset are provided in Table 17. The algorithm of the proposed metric is presented. The time complexity of the algorithm is quadratic and which is consistent with the existing algorithms.

<table>
<thead>
<tr>
<th>Table 17: Statistics of the TUAW Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloggers Count</td>
</tr>
<tr>
<td>Posts Count</td>
</tr>
<tr>
<td>Inlinks Count</td>
</tr>
<tr>
<td>Comments Count</td>
</tr>
<tr>
<td>Weblogs Count</td>
</tr>
<tr>
<td>In-links per post</td>
</tr>
<tr>
<td>Comments per post</td>
</tr>
<tr>
<td>Post per Blogger</td>
</tr>
<tr>
<td>Avg Post Length</td>
</tr>
</tbody>
</table>

**ALGORITHM 3: Identification of the top Influential Bloggers**

**Input:** Data of Posts in web blog  
**Output:** top K Influential bloggers

1. Initialize \(N^b_p, N^b_o, N^b_i, N^b_d, N^p_c, N^p_l\)
2. For each \(b \in B\)
3. FOR each \(p \in P\)
4. \(N^b_p = N^b_p + 1\)
5. \(N^b_d = \text{CalculateActiveDays}(u)\)
6. \(S^b_r = \frac{N^b_p}{(\text{max}(postdate) - \text{min}(postdate))/30}\)
7. \(N^b_c = \text{CalculateComments}(p)\)
8. \(N^b_i = \text{CalculateInlinks}(p)\)
9. \(N^b_o = \text{CalculateOutlinks}(p)\)
10. \(N^b_l = \text{CountChars}(p)\)
11. \(S^b_a = \frac{N^b_l}{N^b_p}\)
12. END FOR
13. END FOR
14. FOR each \(s \in S\)
15. \(N^s_c = \text{CalculateComments}(s)\)
16. \(N^s_i = \text{CalculateInlinks}(s)\)
17. \(N^s_o = \text{CalculateBloggers}(s)\)
18. \(N^s_p = \text{CountPosts}(s)\)

Identification of Influential Bloggers

19. \[ S_{BRank_i}^g = (N_{b_i}^g + N_{p_i}^g + N_{l_i}^g + N_{c_i}^g) \]
20. END FOR
21. ▶ Computation of Modules Score
22. FOR EACH \( b \in B \)
23. \[ S_{Prod}^b = w_pN_{pb}^b + (w_dN_{db}^b + w_{rl}S_{rl}^b) + (w_lN_{lb}^b + w_aS_{ab}^b) \]
24. \[ S_{popu}^b = w_cN_{cb}^b + (w_I N_{ib}^b - w_oN_{ob}^b) \]
25. \[ S_{BRank}^b = \text{getTopBloggers}(S_{BRank}^g) \]
26. ▶ Computation of Post Dialog Feature Set \( (F_{DB}^p) \)
27. \[ S_{infl}^b = w_{pro} S_{prod}^b + w_{popu} S_{popu}^b + w_{bank} S_{BRank}^b \]
28. END FOR
29. STOP ▶ END of Algorithm

4.5.2 Performance Evaluation Measures

MIIB is compared versus the metrics using three performance evaluation measures:

4.5.2.1 OSim

Osim measures the overlapping similarity of outcome of various ranking algorithms. It is the intersection value of the orders normalized by the top records under consideration. For instance, for the given ordered lists of A and B, top k common results are calculated with the help of equation 14:

\[ OSim = \frac{(A \cap B)}{k} \quad (14) \]

4.5.2.2 Spearman's Rank-Order Correlation

It computes the correlation value of the order of the listed results of two ranking algorithms. It analyzes the correlation between MIIB and its modules results. Spearman rank-order correlation is presented as follows in Equation 15:

\[ \text{Spearman Rank Order Correlation} = 1 - 6 \sum k(k^2 - 1) \quad (15) \]

Where \( d \) denotes difference of rank orders which are outcome of the two ordering algorithms and \( n \) denotes items counts being observed. In this case, it is set as 10, 20 and 30.

4.5.2.3 Kendall's Rank Correlation:
It determines the power of dependence in given results of the two ranking algorithms. This is an indirect measure of number of variations in results of two methods. It is represented as $\tau$ and is computed using the equation 8 given as follows:

$$\tau = \frac{(\text{number of concordant pair}) - (\text{number of discordant})}{(1/2)n(n - 1)} \quad (8)$$

4.6 Results and Discussion

The results of this research problem are discussed in four diverse steps. At first, top ten bloggers for each feature are identified which sets up to predict the top bloggers and compare the results of the proposed approaches with existing methods. Next, a comparative analysis of MIIB with the baseline metrics and models presents how each method is different. We use the performance evaluation measures to compare the modules to show their role in finding top bloggers who have impact on other bloggers.

4.6.1 Feature based Evaluation

The top ten bloggers based on each feature are presented in Table 18. S. McNulty enjoys is exclusive as he enjoys top positions in four important features of Activity, Comments, Activeness and BlogRank. E. Sadun is observed to have high ranks by the majority of factors. So, S.McNutty and E.Sadun both are predicted to be among the possible contender for the overall top position. A comparative evaluation of D.Chartier and D. Caolo is noteworthy as both enjoy ranks among first five spots of feature orders. D.Caolo enjoys higher ranks in relatively more number of features and expected to have higher rank as compared to that of D.Chartier. C.Bohon receives the most number of inlinks, but does not enjoy high top rank in any of the other feature. This feature-centric comparative analysis provides a foundation for comparison of MIIB versus the metrics which are regarded as standard methods in the relevant literature.

The Table 18 and Figure 7 present the variation in the rank orders based on each feature. It reveals that S. McNulty has less overall variation than E.Sadun and C. K. Sample III. The cases of D. Charier and D. Caolo are similar, but vary in the case of inlinks received which is an important factor.
Table 18: The top Ten Bloggers using Individual Feature

<table>
<thead>
<tr>
<th></th>
<th>F1-noofposts</th>
<th>F2-noodays</th>
<th>F3-consist</th>
<th>F4-com</th>
<th>F5-inlink</th>
<th>F6-outlink</th>
<th>F7-BRrank</th>
<th>F8-len</th>
<th>F9-avglength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S. McNulty</td>
<td>S. McNulty</td>
<td>B. Dybwad</td>
<td>S. McNulty</td>
<td>C. Bohon</td>
<td>B. Hill</td>
<td>S. McNulty</td>
<td>E. Sadun</td>
<td>C. Ullrich</td>
</tr>
<tr>
<td>2</td>
<td>D. Caolo</td>
<td>D. Caolo</td>
<td>D. Chartier</td>
<td>E. Sadun</td>
<td>E. Sadun</td>
<td>C.K. Sample</td>
<td>E. Sadun</td>
<td>David Chartier</td>
<td>C. Martian</td>
</tr>
<tr>
<td>3</td>
<td>D. Chartier</td>
<td>D. Chartier</td>
<td>S. Bonner</td>
<td>D. Caolo</td>
<td>R. Palmer</td>
<td>M. Scianna mea</td>
<td>D. Caolo</td>
<td>S. McNulty</td>
<td>P. S. Burke</td>
</tr>
<tr>
<td>4</td>
<td>E. Sadun</td>
<td>E. Sadun</td>
<td>C.K. Sample</td>
<td>D. Chartier</td>
<td>D. Caolo</td>
<td>Greg Scher</td>
<td>D. Chartier</td>
<td>D. Caolo</td>
<td>J. Clarke</td>
</tr>
<tr>
<td>5</td>
<td>C.K. Sample</td>
<td>M. Rose</td>
<td>E. Sadun</td>
<td>V. Agreda</td>
<td>M. Schram m</td>
<td>D. Smith</td>
<td>C. Bohon</td>
<td>M. Lu</td>
<td>C. Warren</td>
</tr>
<tr>
<td>6</td>
<td>M. Lu</td>
<td>M. Lu</td>
<td>S. McNulty</td>
<td>M. Lu</td>
<td>M. Rose</td>
<td>D. Touve</td>
<td>M. Lu</td>
<td>M. Rose</td>
<td>B. Terpstra</td>
</tr>
<tr>
<td>7</td>
<td>L.A. Duncan</td>
<td>C. Bohon</td>
<td>D. Caolo</td>
<td>C. Bohon</td>
<td>M. Lu</td>
<td>M. Orchant</td>
<td>V. Agreda</td>
<td>M. Lu</td>
<td>C.K. Sample</td>
</tr>
<tr>
<td>9</td>
<td>M. Rose</td>
<td>M. Schramm</td>
<td>M. Lu</td>
<td>M. Schram m</td>
<td>S. McNul ty</td>
<td>J. Kabili</td>
<td>M. Schramm</td>
<td>C. Bohon</td>
<td>C. Coleman</td>
</tr>
</tbody>
</table>

![Figure 7: The order-alterations of the Influential Bloggers using individual Features](image-url)
4.6.2 Comparison of MIIB and baseline Model

The Table 19 presents the top ten results of modules, iFinder, and MIIB. S.McNutty is at top position by MIIB, however, according to baseline model of iFinder, it is not ranked in top ten even. All the three modules rank him as the top influential blogger. This result is in line with the feature-centric prediction.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Productivity</th>
<th>Popularity</th>
<th>Quality</th>
<th>iFinder</th>
<th>MIIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S. McNulty</td>
<td>S. McNulty</td>
<td>S. McNulty</td>
<td>Cory Bohon</td>
<td>S. McNutty</td>
</tr>
<tr>
<td>2</td>
<td>D. Caolo</td>
<td>E. Sadun</td>
<td>E. Sadun</td>
<td>R. Palmer</td>
<td>E. Sadun</td>
</tr>
<tr>
<td>3</td>
<td>D. Chartier</td>
<td>D. Caolo</td>
<td>D. Caolo</td>
<td>Mat Lu</td>
<td>D. Caolo</td>
</tr>
<tr>
<td>4</td>
<td>E. Sadun</td>
<td>C. Bohon</td>
<td>D. Chartier</td>
<td>C. Warren</td>
<td>D. Chartier</td>
</tr>
<tr>
<td>5</td>
<td>C.K. Sample</td>
<td>D. Chartier</td>
<td>C. Bohon</td>
<td>D. Caolo</td>
<td>C. Bohon</td>
</tr>
<tr>
<td>6</td>
<td>Mat Lu</td>
<td>V. Agreda</td>
<td>V. Agreda</td>
<td>C. Ullrich</td>
<td>V. Agreda</td>
</tr>
<tr>
<td>7</td>
<td>L. A. Duncan</td>
<td>Mat Lu</td>
<td>Mat Lu</td>
<td>S. Sande</td>
<td>Mat Lu</td>
</tr>
<tr>
<td>8</td>
<td>C. Bohon</td>
<td>M.l Rose</td>
<td>M.l Rose</td>
<td>M.l Rose</td>
<td>M.l Rose</td>
</tr>
<tr>
<td>9</td>
<td>M.l Rose</td>
<td>M. Schramm</td>
<td>M. Schramm</td>
<td>V. Agreda</td>
<td>M. Schramm</td>
</tr>
</tbody>
</table>

The baseline ranks C. Bohon as the number one blogger. But, the feature-based analysis displays him 8th in activity, 7th in activeness and is not ranked in top five ranks in any other feature except in-links. This proves that that iFinder offers too much significance to in-links, while MIIB takes all the features into account with respect to their significance. E.Sadun is ranked 2nd by MIIB as expected and is not ranked among top ten bloggers by iFinder. The popularity and BlogRank modules provide her higher ranks. D.Chartier and D.Caolo enjoy 3rd and 4th ranks respectively by MIIB as per anticipation based on feature-wise analysis. It is notable that D.Caolo has higher rank as compared to D.Chartier. MIIB rank C.Bohon at 5th rank as expected mainly due to its similar positions in feature-wise results. As the iFinder and MIIB results are quite different from one another, thus performance evaluation measure may not be applied for comparison.

4.6.3 Comparison of MIIB vs Existing Metrics

Table 20 presents the top ten bloggers using for comparison of MIIB against the metrics of MEIBI and MEIBIX [62]. As the bloggers are common but variations exist so the results of performance evaluation measures are displayed in Table 21. The high OSim
represents higher common bloggers. Low correlation reveals variations in the results of the proposed method versus those of the existing metrics. We, compare the three top bloggers. MEIBI and MEIBIX rank C. Bohon as the first position older among all bloggers. Feature based analysis reveal that he enjoys top position only in in-links feature and is not ranked in even top five ranks by any other feature. MIIB provides him the 5th position. R. Palmer is ranked 8th in the consistency and 3rd in inlinks. Baseline metrics rank him 2nd while the MIIB rank him 10th. S. Sande need attention for his ranks due to the fact that he enjoy 8th position in in-links and does not enjoy any other rank by any factor, but MEIBI and MEIBIX rank him at 3rd position. It proves that the existing metrics give extra significance to inlinks. The relevant literature [62, 66] also argue that an in-link may not necessarily be in favor as it can be against the start post’s point of view and thus giving it extra significance may be improper.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C. Bohon</td>
<td>C. Bohon</td>
<td>S. McNutty</td>
</tr>
<tr>
<td>2</td>
<td>R. Palmer</td>
<td>R. Palmer</td>
<td>E. Sadun</td>
</tr>
<tr>
<td>3</td>
<td>Steven Sande</td>
<td>Steven Sande</td>
<td>D. Caolo</td>
</tr>
<tr>
<td>4</td>
<td>E. Sadun</td>
<td>E. Sadun</td>
<td>D. Chartier</td>
</tr>
<tr>
<td>5</td>
<td>M.1 Rose</td>
<td>C. Warren</td>
<td>C. Bohon</td>
</tr>
<tr>
<td>6</td>
<td>M. Schramm</td>
<td>M.1 Rose</td>
<td>V. Agreda</td>
</tr>
<tr>
<td>7</td>
<td>C. Warren</td>
<td>M. Schramm</td>
<td>Mat Lu</td>
</tr>
<tr>
<td>8</td>
<td>D. Caolo</td>
<td>Mat Lu</td>
<td>M.1 Rose</td>
</tr>
<tr>
<td>9</td>
<td>Mat Lu</td>
<td>D. Caolo</td>
<td>M. Schramm</td>
</tr>
<tr>
<td>10</td>
<td>Brett Terpstra</td>
<td>Brett Terpstra</td>
<td>R. Palmer</td>
</tr>
</tbody>
</table>

**Table 20**: Top Ten bloggers based on Modules, iFinder and the MIIB

**Table 21**: Evaluation of MIIB vs Baseline Metrics

<table>
<thead>
<tr>
<th></th>
<th>OSim</th>
<th>Spearman Correlation</th>
<th>Kendall Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEIBI vs MEIBIX</td>
<td>1</td>
<td>0.9515</td>
<td>0.8667</td>
</tr>
<tr>
<td>MEIBI vs MIIB</td>
<td>0.8</td>
<td>0.2242</td>
<td>0.2</td>
</tr>
<tr>
<td>MEIBIX vs MIIB</td>
<td>0.8</td>
<td>0.22</td>
<td>0.16</td>
</tr>
</tbody>
</table>

### 4.6.3 Module-wise Evaluation

Figure 8 presents a comparison of MIIB modules. A careful observation proves that bloggers’ overall positions based on each module are fairly consistent. The popularity is a direct measure of a blogger’s impact and its top results are closely similar to MIIB outcome. In case of productivity, variation is visible. Please note that the productivity is
an indirect measure of a blogger’s influence. So it is important to consider not only activity feature but also other features considered in the productivity module as only considering the post creating capability is not a true measure of a blogger’s impact. The module-based comparison is provided in Figure 9.

![Figure 8: A Module-wise Comparative Analysis](image)

The Figure 9 validates the above cited analysis and proves that all the modules of MIIB play their role to find top bloggers in the community.

![Figure 9: Comparison of MIIB and its Modules](image)
4.6.4 Evaluation using Performance Evaluation Measures

The existing relevant literature lacks to compare the methods using evaluation measures. We introduce the use of three performance evaluation measures and their results are elaborated. The comparative analysis has been carried out using the values of k as 10, 20, 30 and then all the bloggers are considered as results provided in Tables 22, 23 and 24.

<table>
<thead>
<tr>
<th>Comparison between</th>
<th>Dataset</th>
<th>Top 30</th>
<th>Top 20</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productivity vs Popularity</td>
<td>0.8842</td>
<td>0.8641</td>
<td>0.8316</td>
<td>0.8362</td>
</tr>
<tr>
<td>Productivity vs BlogRank</td>
<td>0.8962</td>
<td>0.8784</td>
<td>0.8502</td>
<td>0.8625</td>
</tr>
<tr>
<td>Popularity vs BlogRank</td>
<td>0.9997</td>
<td>0.9996</td>
<td>0.9994</td>
<td>0.9988</td>
</tr>
<tr>
<td>Productivity vs MIIB</td>
<td>0.8962</td>
<td>0.8784</td>
<td>0.8502</td>
<td>0.8625</td>
</tr>
<tr>
<td>Popularity vs MIIB</td>
<td>0.9997</td>
<td>0.9996</td>
<td>0.9994</td>
<td>0.9988</td>
</tr>
<tr>
<td>BlogRank vs MIIB</td>
<td>0.9978</td>
<td>0.9988</td>
<td>0.9898</td>
<td>0.9786</td>
</tr>
</tbody>
</table>

Spearman rank-order correlation finds the correlation coefficient between the results of MIIB and its modules. It reveals that Popularity and BlogRank modules have higher correlations with the MIIB in comparison of the module of Productivity. The value of correlation alters as the value of k increases, but overall there is a high correlation values for all the k values.

<table>
<thead>
<tr>
<th>Comparison between</th>
<th>Top 30</th>
<th>Top 20</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productivity vs Popularity</td>
<td>0.70732</td>
<td>0.61839</td>
<td>0.62105</td>
</tr>
<tr>
<td>Productivity vs BlogRank</td>
<td>0.73415</td>
<td>0.65057</td>
<td>0.63158</td>
</tr>
<tr>
<td>Popularity vs BlogRank</td>
<td>0.95406</td>
<td>0.96782</td>
<td>0.98947</td>
</tr>
<tr>
<td>Productivity vs MIIB</td>
<td>0.73415</td>
<td>0.65057</td>
<td>0.63158</td>
</tr>
<tr>
<td>Popularity vs MIIB</td>
<td>0.95406</td>
<td>0.96782</td>
<td>0.98947</td>
</tr>
<tr>
<td>BlogRank vs MIIB</td>
<td>0.96880</td>
<td>0.97989</td>
<td>0.98766</td>
</tr>
</tbody>
</table>

Kendall correlation measures the power of correlation between the orders of MIIB and its modules. It indirectly shows the variations in the results of ranking orders of the different methods. In this case, the higher values show the less number of variations.

<table>
<thead>
<tr>
<th>Comparison between</th>
<th>Top 30</th>
<th>Top 20</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productivity vs Popularity</td>
<td>0.93333</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Productivity vs BlogRank</td>
<td>0.93333</td>
<td>0.85</td>
<td>0.8</td>
</tr>
<tr>
<td>Popularity vs BlogRank</td>
<td>1</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>Productivity vs MIIB</td>
<td>0.93333</td>
<td>0.85</td>
<td>0.8</td>
</tr>
<tr>
<td>Popularity vs MIIB</td>
<td>1</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>BlogRank vs MIIB</td>
<td>0.95</td>
<td>0.95</td>
<td>1</td>
</tr>
</tbody>
</table>
OSim finds the overlapping results of the two methods. In this case, it counts the common number of bloggers among top $k$ ranks. For the entire dataset, Osim is 1. Popularity and BlogRank have high common results as those of the MIIB. All the MIIB modules have very high correlated values for top k values, which confirm that all the modules play their role to find the influential bloggers.

4.7 MIIB Evaluation using Engadget and Techcrunch Dataset

We have discussed the detailed analysis of our proposed metric MIIB on the dataset of TUAW, which is widely used in the relevant literature. For comparison with recent datasets to analyze how the proposed metric perform with respect to other existing metrics, we selected Engadget and Techcrunch datasets.

4.7.1 Engadget Dataset

The Engadget blog\footnote{http://www.engadget.com/}. Accessed April 08,2016. covers electronic gadget and shares the news about the technology topics where common users can discuss on technology topics. Engadget is regarded as one of the top blogs as it has been awarded as the best Tech Blog for two years by the Bloggie and it has been ranked among top five blogs in the world as declared by Time magazine in 2010\footnote{https://en.wikipedia.org/wiki/Engadget}. According to Wikipedia\footnote{https://en.wikipedia.org/wiki/Engadget}. Accessed April 09, 2016., in year 2013, Engadget has been ranked in top five blogs in the Blogosphere as ranked by the Technorati, which is the world renowned publisher advertising platform and provides web related solutions to all over the world.

Engadget dataset is available free of cost\footnote{http://users.sch.gr/lakritid/code.php?c=3}. Accessed April 10, 2016. and has been used in the recent relevant research works [66, 68, 69] and used in evaluation of BI-index, BP-index as well as the other metrics which are derived from these two metrics. The characteristics of the dataset are given in the Table 25.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Engadget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloggers</td>
<td>93</td>
</tr>
</tbody>
</table>

Identification of Influential Bloggers

<table>
<thead>
<tr>
<th></th>
<th>Blog Posts</th>
<th>In-links</th>
<th>Comments</th>
<th>Blog Post per Blogger</th>
<th>In-links per Blog Post</th>
<th>Comments per Blog Post</th>
<th>Average Length per Blog Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63,358</td>
<td>319,880</td>
<td>3,672,819</td>
<td>681</td>
<td>5</td>
<td>58</td>
<td>180</td>
</tr>
</tbody>
</table>

4.7.2 Comparison of MIIB using Engadget with Baseline metrics
MIIB has been compared with the recent metrics of BI-Index and BP-Index and the results are provided in Table 26.

A careful analysis of the table depicts that the top rank of blogger Murph D. is consistent in the baseline as well as in module and MIIB which verifies that the MIIB and its modules find the top blogger. But the case of Ziegler C. is noteworthy. Ziegler is ranked very high by the baseline methods but the proposed methods do not rank him highly. As in single feature-wise analysis, he is ranked 8th in activity, 8th in inlinks, and 9th in comments which are the two important features to measure a blogger’s influence.

Table 26: A Comparison of MIIB, Modules versus Baselines

<table>
<thead>
<tr>
<th></th>
<th>BI-Index</th>
<th>BP-Index</th>
<th>Productivity module</th>
<th>Popularity module</th>
<th>BlogRank module</th>
<th>MIIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Murph D.</td>
<td>Murph D.</td>
<td>Murph D.</td>
<td>June L.</td>
<td>Murph D.</td>
<td>Murph D.</td>
</tr>
<tr>
<td>2</td>
<td>Ziegler C.</td>
<td>Ziegler C.</td>
<td>Block R.</td>
<td>Murph D.</td>
<td>Miller P.</td>
<td>Miller P.</td>
</tr>
<tr>
<td>3</td>
<td>Savov V.</td>
<td>Savov V.</td>
<td>Miller P.</td>
<td>Ricker T.</td>
<td>Ricker T.</td>
<td>Ricker T.</td>
</tr>
<tr>
<td>4</td>
<td>Miller P.</td>
<td>Ricker T.</td>
<td>Ricker T.</td>
<td>Melanson D.</td>
<td>Block R.</td>
<td>Block R.</td>
</tr>
<tr>
<td>5</td>
<td>Flatley J.</td>
<td>Topolsky J.</td>
<td>Block R.</td>
<td>Patel N.</td>
<td>June L.</td>
<td>June L.</td>
</tr>
<tr>
<td>6</td>
<td>Stevens T.</td>
<td>Miller P.</td>
<td>Ricker T.</td>
<td>Topolsky J.</td>
<td>Melanson D.</td>
<td>Melanson D.</td>
</tr>
<tr>
<td>7</td>
<td>Stern J.</td>
<td>Stevens T.</td>
<td>Patel N.</td>
<td>Block R.</td>
<td>Patel N.</td>
<td>Patel N.</td>
</tr>
<tr>
<td>8</td>
<td>Ricker T.</td>
<td>Flatley J.</td>
<td>Topolsky J.</td>
<td>Melanson D.</td>
<td>Rojas P.</td>
<td>Topolsky J.</td>
</tr>
<tr>
<td>9</td>
<td>Topolsky J.</td>
<td>Miller R.</td>
<td>Ziegler C.</td>
<td>Ziegler C.</td>
<td>Topolsky J.</td>
<td>Ziegler C.</td>
</tr>
<tr>
<td>10</td>
<td>Miller R.</td>
<td>Stern J.</td>
<td>Blass E.</td>
<td>Miller R.</td>
<td>Ziegler C.</td>
<td>Rjas P.</td>
</tr>
</tbody>
</table>

The comparison of Miller P. and Ricker T. is noticeable as both enjoy as in feature-wise analysis, both enjoy ranks in top five positions but Miller P. enjoys relatively better rank positions in the ordering in the features of inlinks, activity, comments. The baselines rank Ricker T. higher than Miller P. while the modules of productivity and blog rank as well
as MIIB ranks Miller P. as compare to Ricker T. which is proper ordering. This individual blogger analysis reveals that the proposed metric MIIB and its modules identify the top bloggers and gives them proper ranks.

Table 27: A comparison between MIBSA and baseline metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>OSim</th>
<th>Spearman’s Correlation</th>
<th>Kendall’s Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productivity vs. MIIB</td>
<td>0.90</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>Popularity vs. MIIB</td>
<td>0.90</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>BlogRank vs. MIIB</td>
<td>0.90</td>
<td>0.87</td>
<td>0.71</td>
</tr>
<tr>
<td>BP Index vs. MIIB</td>
<td>0.60</td>
<td>0.55</td>
<td>0.37</td>
</tr>
<tr>
<td>BI Index vs. MIIB</td>
<td>0.50</td>
<td>0.45</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 27 presents the comparative analysis of the proposed method, its modules and the baselines using the performance evaluation measure. Osim measures the overlapping similarity of the two orders. The high values of overlapping similarity suggest that overall bloggers identification is similar.

The rank-order correlation finds the correlation between the two orders whereas the Kendall correlation finds the alterations in ordering of the two rankings. The correlation results are relatively different which suggest that the order of the ranks of the methods vary which is understandable as the baselines uses three main features of activity, in-links and comments along with temporal factor while the proposed metric uses as many as nine different features. The overall similarity of modules is very high as compared to the baselines.

4.7.3 Techcrunch Dataset

Techcrunch\textsuperscript{16} is a world renowned web log for sharing news and articles related to information technology companies and business organizations. According to Wikipedia\textsuperscript{17}, as many as about two million users access the blog every month which reveals the volume of the data of the blog.

\textsuperscript{16} \url{http://techcrunch.com/}. Accessed April 09, 2016.
\textsuperscript{17} \url{https://en.wikipedia.org/wiki/TechCrunch}. Accessed April 09, 2016.
Techcrunch dataset containing features of bloggers is available free of cost\textsuperscript{18} and has been used in the recent relevant papers [66]. The characteristics of the dataset are presented in Table 28.

Table 28: Techcrunch Dataset Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Engadget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloggers</td>
<td>107</td>
</tr>
<tr>
<td>Bog Posts</td>
<td>19,464</td>
</tr>
<tr>
<td>In-links</td>
<td>193,808</td>
</tr>
<tr>
<td>Comments</td>
<td>746,561</td>
</tr>
<tr>
<td>Blog Post per Blogger</td>
<td>181</td>
</tr>
<tr>
<td>In-links per Blog Post</td>
<td>10</td>
</tr>
<tr>
<td>Comments per Blog Post</td>
<td>38</td>
</tr>
<tr>
<td>Average Length per Blog Post</td>
<td>169</td>
</tr>
</tbody>
</table>

4.7.4 Comparison of MIIB using Techcrunch data with existing metrics

Table 29 presents the top ten bloggers ranked by the baseline, MIIB and its modules. A careful analysis reveals that the top bloggers in the Techcrunch blog. s. The accumulative value of three modules of popularity, BlogRank and productivity scores provides the MIIB score of the bloggers.

Table 29: Blogger's Ranking based on MIIB, Modules and Baseline

<table>
<thead>
<tr>
<th>H-index</th>
<th>Productivity module</th>
<th>Popularity module</th>
<th>BlogRank</th>
<th>MIIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Michael A.</td>
<td>Michael A.</td>
<td>Michael A.</td>
<td>Michael A.</td>
</tr>
<tr>
<td>2</td>
<td>Erick S.</td>
<td>Erick S.</td>
<td>MG Siegler</td>
<td>Erick S.</td>
</tr>
<tr>
<td>3</td>
<td>MG Siegler</td>
<td>Jason K.</td>
<td>Erick S.</td>
<td>Jason K.</td>
</tr>
<tr>
<td>4</td>
<td>Duncan R.</td>
<td>MG Siegler</td>
<td>Jason K.</td>
<td>Robin W.</td>
</tr>
<tr>
<td>5</td>
<td>Jason K.</td>
<td>Robin W.</td>
<td>Robin W.</td>
<td>Duncan R.</td>
</tr>
<tr>
<td>6</td>
<td>Mark H.</td>
<td>Leena R.</td>
<td>John B.</td>
<td>Leena R.</td>
</tr>
<tr>
<td>7</td>
<td>Robin W.</td>
<td>Duncan R.</td>
<td>Guest A.</td>
<td>Guest A.</td>
</tr>
<tr>
<td>8</td>
<td>Leena R.</td>
<td>Mark H.</td>
<td>Leena R.</td>
<td>MG Siegler</td>
</tr>
<tr>
<td>9</td>
<td>Marshall K.</td>
<td>Guest A.</td>
<td>Mark H.</td>
<td>John B.</td>
</tr>
<tr>
<td>10</td>
<td>Guest A.</td>
<td>John B.</td>
<td>Sarah L.</td>
<td>Mark H.</td>
</tr>
</tbody>
</table>

The overall common bloggers’ count is significantly high suggesting the proper approach by the proposed as well as a baseline. Kendall similarity measure depicts the alterations in ordering in the compared lists and higher values shows that the ordering by the baseline and compared results is high whereas the correlation in the results is higher as well.

MIIB has been compared using the three widely used datasets of TUAW, Engadget, and Techcrunch and the results are compared with state of the art existing methods in the relevant literature.

4.8 Chapter Summary
In this chapter, a modular approach presents MIIB, a new metric to find influential bloggers in the blogosphere. In addition to important existing factors, more features are introduced. Top blog sites and later top bloggers are identified in those blogs. The three modules of popularity, productivity and BlogRank are compared using the evaluation measures. It has been found that for productivity measure, in addition to activity, consistency and activeness are also important so that a blogger should maintain his/her influence in a community on consistently via regular activity. For popularity, our findings are similar to existing work that both inlinks and comments are the appropriate factors for computing the impact of a blogger. However, the existing work give too much importance to inlinks whereas the proposed metric provides due importance to both the factors as both these factors, in perspective of the social web domain, can be in favor as well against. The importance of blog is another new factor introduced has been found very effective. The proposed metric evaluated against the existing metrics and models is found effective in identification of top influential bloggers.
Chapter 5: Mixed-Opinion Classification using Thread Features
Chapter 5

Mixed-Opinion Classification using Thread Features

5.1 Introduction
As discussed earlier, the social web forums contain content about various topics. Opinions and emotions are part of online content as its topics are near to our daily life. In opinion mining, sentiment polarity categorizes whether sentiment is positive or negative. Similarly, emotion arousal determines the strength of the emotion. In this work, we identify mixed-opinion posts and threads i.e., having positive and negative sentiments.

5.2 Why Classifying Content for Mixed-Opinion Identification
Identification of such discussions has various potential applications such as monitoring public views, making products related business decisions and predicting users’ behaviors. Sentiment mixture is helpful to find the social and political issues about which the people express bifurcated sentiments, to find out certain products which receive reviews of diverse emotional valence from users and to foresee reaction of forthcoming government policies. We intend to explore the characteristics of mixed-opinion public views. A study finds contentions from discussions and debates [83], to mine contrasting opinion on political texts [82], to discover sentiment based contradictions [88]. Mixed opinion analysis help to find the topics for which diverse opinions have been expressed, to understand mixed-opinions, and to analyze their evolution over time [14]. Detection of antagonistic or over-heated language in communication such as email is another application of identification of mixed opinion content.

5.3 Problem Formulation and Problem Statement
Let us formally formulate and state the problem.
Mixed-Opinion Classification using Thread Features

5.3.1 Problem Formulation

A user can initiate a new topic by creating a new thread in an online forum. The initial post sets the topic for discussion. Our aim is to identify only those posts and threads which have high positive and high negative emotions in it. During discussion, topic can drift in the threads and our assumption may not hold right always but such exceptional cases are out of the scope of this research.

Formally, we define a forum post \( p \) is a words order in a Vocabulary set \( V \), a forum thread \( t \) is a sequence of posts i.e., \( t = \{ p_1, p_2, ..., p_L \} \) where \( p_i \) is the \( i^{th} \) post in the thread and forum \( f \) to be a collection of threads \( f = \{ t_1, t_2, ..., t_m \} \) where \( t_j \) is a thread. A thread is initiated by a user \( u \in U \), and \( U = \{ u_1, u_2, ..., u_n \} \) where \( n \) is the number of forum users. A post having high positive and negative sentiment values is regarded as mixed-sentiment post, denoted as \( p_{ms} \) and denoted as \( p_{nms} \) otherwise. Similarly, a thread containing set of positive and negative posts is taken as mixed-sentiment thread, denoted as \( t_{ms} \) and denoted as \( t_{nms} \).

5.3.2 Problem Statement

Given an online forum \( f \) and the set of thread \( t \), our aim is to classify each post \( p_i \) and thread \( t_j \) into one of the two given classes: Mixed-opinion posts (denoted by \( p_{ms} \)) or otherwise (\( p_{nms} \)) in case of posts and Mixed-opinion Thread (denoted by \( t_{ms} \)) or otherwise (\( t_{nms} \)) in case of threads.

5.4 Feature Engineering

Let us propose thread-structure feature set in addition to various proposed dialog and sentiment based features. The purpose is to investigate the effects of the proposed features in classification of mixed-sentiment posts and threads. The Table 30 provides the list of symbols used to calculate the proposed features and the post and the thread features are given in Tables 31 and 32 respectively.
Mixed-Opinion Classification using Thread Features

**Table 30**: List of Symbols used for Mixed-Opinion Classification

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>Set of Threads</td>
</tr>
<tr>
<td>( P )</td>
<td>Set of Posts</td>
</tr>
<tr>
<td>( U )</td>
<td>Set of Users</td>
</tr>
<tr>
<td>( t )</td>
<td>( t \in T )</td>
</tr>
<tr>
<td>( p )</td>
<td>( p \in P )</td>
</tr>
<tr>
<td>( u )</td>
<td>( u \in U )</td>
</tr>
<tr>
<td>( N_p^t )</td>
<td>Posts in a thread</td>
</tr>
<tr>
<td>( N_u^t )</td>
<td>Users in a thread</td>
</tr>
<tr>
<td>( N_{\text{un}}^p )</td>
<td>Number of Username mentioned in post</td>
</tr>
<tr>
<td>( N_q^p )</td>
<td>Number of textQuoted in post</td>
</tr>
<tr>
<td>( N_{\text{ur}}^p )</td>
<td>Number of URL in post</td>
</tr>
<tr>
<td>( N_c^p )</td>
<td>Number of Capital words in post</td>
</tr>
<tr>
<td>( N_s^p )</td>
<td>Sentiment words in post (using SentiWordNet)</td>
</tr>
<tr>
<td>( N_{pw}^p ) or ( N_{nmw}^p )</td>
<td>Words in a post / thread</td>
</tr>
<tr>
<td>( S_p^p ) or ( S_n^p )</td>
<td>The Positive score of the post/thread (using SentiWordNet)</td>
</tr>
<tr>
<td>( S_{np}^p ) or ( S_{nt}^p )</td>
<td>The Negative score of the post/thread (using SentiWordNet)</td>
</tr>
<tr>
<td>( N_{np}^p ) or ( N_{nw}^p )</td>
<td>Number of positive words in post/thread. (A word is a positive word if its positive value &gt; 0 and negative value = 0 ) (using SentiWordNet)</td>
</tr>
<tr>
<td>( N_{nw}^p ) or ( N_{nw}^t )</td>
<td>Number of negative words in post. (A word is a positive word if its negative &gt; 0 and positive value = 0 ) (using SentiWordNet)</td>
</tr>
<tr>
<td>( N_{mw}^p ) or ( N_{mw}^t )</td>
<td>Number of Mixed-opinion words in post/thread. (A word is a Mixed-word if its negative &gt; 0 and positive value &gt; 0 ) (using SentiWordNet)</td>
</tr>
<tr>
<td>( S_{npw}^p ) or ( S_{npw}^t )</td>
<td>Mixed-opinion Score of the post based on Mixed-words.</td>
</tr>
</tbody>
</table>

5.4.1 Post Features

The post features are further categorized into Sentiment and Dialog features.

5.4.1.1 Sentiment Features (Lexical)

The sentiment feature set takes into account a user’ sentiment. The features are computed using various resources like sentiment lexicon (e.g., SentiWordNet [30] and WordNet-Affect) as well as sentiment analysis tool (e.g., SentiStrength [105, 106] and LIWC [107]). SentiWordNet has been used for computation of sentiment features. The pPositiveScore and pNegativeScore denotes positive and negative sentiment scores of a post respectively and their absolute difference shows mixed nature of both sentiment score (pSentiScore). Similarly the sentiment word feature (pSentiWordsScore) is the difference of the positive and negative words counts. The difference depicts mixed-opinion nature topic discussed in the content. The polarity of a word is identified using SentiWordNet. A positive word has positive value greater than zero and the negative value is equal to zero. Similarly, a negative word has negative value greater than zero.
while positive value is equal to zero. The Mixed-word has both positive and negative values greater than zero (pMixWordsScore). An opinionative post is usually lengthier than an informative one [114] so the post length (pPostLength) is considered as a feature.

5.4.1.2 Dialog Features

Let us assume that Mixed-opinion topics have a higher chance of dialog and thus the discussions of mixed-opinion topics have higher chance of conversations among the users. Let us posit that dialog features help to detect mixed-opinion posts. The first feature is the existence of URLs (boolURL) as the user share links to the web pages or other posts within the forum as an argument or evidence in favor of their point of view. A user mentions another user’s name who has already posted his/her comments to seek his/her attention for replying his question or sharing a similar point of view as that of the user. Such user mentions may be common in an emotional conversation as compared to technical discussion (boolUsername). Likewise, a user copies post text of an earlier post (boolQuotedText). The content in upper case depicts intensifier or shouting [37] (boolCapital) is helpful to recognize dialog. Boolean features exhibit the existence of certain characteristics and the similar features consider count at thread-level.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Sentiment Features</strong></td>
<td></td>
</tr>
<tr>
<td>$S_p$</td>
<td>pSentiScore</td>
<td>Sentiment Score of the post</td>
</tr>
<tr>
<td>$S_w$</td>
<td>pSentiWords</td>
<td>Sentiment Score based on Sentiment Words</td>
</tr>
<tr>
<td>$S_d$</td>
<td>pMixWordsScore</td>
<td>Mixed Words Score based on Mixed-Opinion Words</td>
</tr>
<tr>
<td>$N_l$</td>
<td>pPostLength</td>
<td>Post length</td>
</tr>
<tr>
<td></td>
<td><strong>Dialog Features</strong></td>
<td></td>
</tr>
<tr>
<td>$b_{un}$</td>
<td>boolUsername</td>
<td>Existence of Username mentioned in the post</td>
</tr>
<tr>
<td>$b_{qu}$</td>
<td>boolQuotedText</td>
<td>Existence of earlier thread posts quoted in the post</td>
</tr>
<tr>
<td>$b_{ur}$</td>
<td>boolURL</td>
<td>Existence of URL in post</td>
</tr>
<tr>
<td>$b_{uc}$</td>
<td>boolCapital</td>
<td>Existence of Capital Case words in the post</td>
</tr>
</tbody>
</table>
5.4.2 Thread Features

Thread features are of three categories: Sentiment, Dialog and Structural. The sentiment and dialog features are similar to post features. The structural features are not limited to a particular language and do not require a lexicon. The features are represented using formal notations and are discussed using feature name for understanding. A thread feature value is based on all the posts in it.

5.4.2.1 Sentiment Features

The thread sentiment features consider the accumulative scores of all the respective posts sentiment features. The thread sentiment score (tSentiScore) is the absolute difference of the accumulative positive score and that of negative score of all the posts. Similarly, the sentiment words score (tSentiWordsScore) is the absolute difference of the accumulative scores of positive words and negative words in all the posts of the thread. The difference highlights the mixed-opinion nature of the nature. In both features, the calculated differences show that the smaller the difference, the higher the chance that the discussion is having mixed-opinion views. The third feature (tSentiDivScore) is the number of the sentiment diverse words normalized by the word count in the thread.

5.4.2.2 Dialog Features

These features detect dialog on the same notion as already defined in post features. The first feature is the number of URLs (numURL) in the thread as users emphasize their point of view by sharing links to other similar topic threads within the forum or to web pages. A user mentions the user of earlier post in the thread to address them (numUsername) or shares post from earlier posts as a reply to that earlier post (numQuotedText) respectively. The content in upper case represents shouting (numCapital). We introduce discrete features anticipating that there is higher chance of multiple presences in threads and anticipate higher values of the above-cited features help to identify mixed-opinion threads.
5.4.2.3 Thread-Structure Features

Let us suppose that a mixed-opinion discussion is likely to possess certain characteristics. For instance, it may contain relatively more number of posts (numPosts), consists of lengthier text (numThreadWords), to present emotional views over a long period (numThreadDuration) as compared to a factual or less opinionative thread. A user may share consecutive posts to express more evidences to strength their aforementioned opinionative point of views (numConsPosts). A user may share an opinionative post; another user may reply it and the first user may again post his comment forming a post-reply cycle (numCycleUSers).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S^t_s )</td>
<td>tSentiScore</td>
<td>Sentiment Score of the thread</td>
</tr>
<tr>
<td>( S^t_w )</td>
<td>tSentiWordsScore</td>
<td>Sentiment Score of the thread based on Sentiment Words</td>
</tr>
<tr>
<td>( S^t_{mw} )</td>
<td>tMixWordsScore</td>
<td>Diversity Score of the thread based on Contradictory Words</td>
</tr>
</tbody>
</table>

### Dialog Features

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N^t_{un} )</td>
<td>numUsername</td>
<td>Number of Username mentioned in the thread</td>
</tr>
<tr>
<td>( N^t_q )</td>
<td>numQuotedText</td>
<td>Number of quoted-text posts in a thread</td>
</tr>
<tr>
<td>( N^t_{ur} )</td>
<td>numURL</td>
<td>Number of URL given in thread</td>
</tr>
<tr>
<td>( N^t_c )</td>
<td>numCapital</td>
<td>Number of Capital Case words</td>
</tr>
</tbody>
</table>

### Thread-Structure Features

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N^t_{cp} )</td>
<td>numConsPosts</td>
<td>Consecutive Posts by the user</td>
</tr>
<tr>
<td>( N^t_{cu} )</td>
<td>numConvCycles</td>
<td>Number of Conversation/Dialog Cycles of users exist in the thread</td>
</tr>
<tr>
<td>( N^t_p )</td>
<td>numPosts</td>
<td>Posts in a thread</td>
</tr>
<tr>
<td>( N^t_c )</td>
<td>numThreadWords</td>
<td>Thread Characters Count</td>
</tr>
<tr>
<td>( N^t_d )</td>
<td>numThreadDuration</td>
<td>Number of days of a thread</td>
</tr>
</tbody>
</table>

The algorithms and the proposed framework for mixed-opinion classification of posts and threads in the social web are presented in Algorithm 4, Algorithm 5 and Figure 10 respectively. The time complexity of both the proposed algorithms is quadratic and which is consistent with the existing algorithms.
**Algorithm 4:** Mixed-Opinion Classification of Posts in Online Forum

**Input:** Data of Posts in Online Forum

**Output:** Post classified as Mixed-Opinion or non-Mixed-opinion

1. Initialize $S^p_s$, $S^p_p$, $S^p_n$, $N^p_p$, $N^p_m$, $N^p_w$.
2. FOR each $t \in T$
3. FOR each $p \in P$
4. $N^p_w = \text{CountWords}(p)$
5. $S^p_p = \text{CalculatePositiveSWNScore}(p)$
6. $S^p_n = \text{CalculateNegativeSWNScore}(p)$
7. $N^p_{pw} = \text{CountPositiveSWNWords}(p)$
8. $N^p_{nw} = \text{CountNegativeSWNWords}(p)$
9. $N^p_{mw} = \text{ComputeMixedOpinionSWNWords}(p)$
10. **Computation of Sentiment Feature Set** ($F^p_s$)
11. $S^p_s = S^p_p - S^p_n$
12. $S^p_w = N^p_{pw} - N^p_{nw}$
13. $S^p_{mw} = \frac{N^p_{mw}}{N^p_w}$
14. $N^p_l = \text{ComputePostLength}(p)$
15. $F^p_s = [F^p_s, S^p_s, S^p_p, S^p_n, N^p_l]$ (Computation of Post Dialog Feature Set ($F^p_d$))
16. IF $p$ contains URL THEN $b^p_{ur} = 1$ END IF
17. IF $p$ contains Capital Word THEN $N^p_c = N^p_c + 1$ END IF
18. IF $p$ contains Username of Earlier posts in the thread $t$ THEN $b^p_{un} = 1$ END IF
19. IF $p$ contains Quoted text from Earlier posts in the thread $t$ THEN $b^p_q = 1$ END IF
20. $F^p_d = [N^p_{ur}, N^p_c, b^p_{un}, b^p_q]$ (End of the Algorithm)
21. END FOR
22. END FOR
23. Class = Classifier($F^p_s$, $F^p_d$)
24. IF Class = 1 then $p' = p_m$ ELSE $p' = p_{nm}$ END IF
25. STOP
**Mixed-Opinion Classification using Thread Features**

**ALGORITHM 5:** Mixed-Opinion Classification of Threads in Online Forum

**Input:** Data of Threads in Online Forum

**Output:** Thread classified as subjective or non-Subjective

1. Initialize $N_{cp}^t$, $N_{cu}^t$, $N_u^t$, $N_d^t$, $N_c^t$, $N_p^t$
2. FOR each $t \in T$
3. FOR each $p \in P$
4. $N_w^p = \text{CountWords}(p)$
5. $N_{pw}^p = \text{CountPositiveSWNWords}(p)$
6. $N_{nw}^p = \text{CountNegativeSWNWords}(p)$
7. $S_{pw}^p = \text{ComputePositiveSWNScore}(p)$
8. $S_{nw}^p = \text{ComputeNegativeSWNScore}(p)$
9. END FOR
10. END FOR
11. $S_{p}^t = \frac{S_{pw}^t}{N_T}$
12. $S_{n}^t = \frac{S_{nw}^t}{N_T}$
13. $N_{pw}^t = \sum_{i=1}^{N_p^t} N_{pw}^p$
14. $N_{nw}^t = \sum_{i=1}^{N_p^t} N_{nw}^p$
15. $N_{mw}^t = \sum_{i=1}^{N_p^t} N_{mw}^p$
16. **Computation of Thread Sentiment Feature Set ($F_s^t$)**
17. $S_{sw}^t = N_{pw}^t - N_{nw}^t$
18. $S_{sw}^t = N_{nw}^t - N_{pw}^t$
19. $S_{nw}^t = N_{nw}^t / N_w^t$
20. **Computation of Thread Dialog Feature Set ($F_d^t$)**
21. $N_{un}^t = \sum_{i=1}^{N_p^t} N_{un}^i$
22. $N_{d}^t = \sum_{i=1}^{N_p^t} N_{d}^i$
23. $N_{ur}^t = \sum_{i=1}^{N_p^t} N_{ur}^i$
24. $N_{ur}^t = \sum_{i=1}^{N_p^t} N_{ur}^i$
25. $N_c^t = \sum_{i=1}^{N_p^t} N_c^i$
26. **Computation of Thread Structure Feature Set ($F_b^t$)**
27. $N_p^t = N_p^t + 1$
28. IF $u_{i-1}^p = u_{i+1}^p$ THEN
29. $N_{cp}^t = N_{cp}^t + 1$
30. End IF
31. IF $u_{i-1}^p = u_{i+1}^p$ THEN
32. $N_{cu}^t = N_{cu}^t + 1$
33. End IF
34. $N_{w}^t = \sum_{i=1}^{N_p^t} N_{w}^i$
35. $N_{d}^t = \text{CountNumOfDays}(t)$
36. $F_d^t = [F_d^p; N_{cp}^t, N_{cu}^t, N_u^t, N_d^t, N_c^t, N_p^t, N_w^t]$
37. END FOR
38. $\text{Class} = \text{Classifier}(F_d^p, F_b^p)$
39. IF $\text{Class} = 1$ THEN
40. \hspace{1em} $p' = p_s$
41. Else
42. \hspace{1em} $p' = p_{ns}$
43. End IF
44. STOP ▷ End of the Algorithm

![Diagram](image-url)

**Figure 10**: The proposed framework for mixed-opinion classification of Online Forum

5.5 Experimental Setup

Let us describe the classification algorithms applied, the BBC forum dataset [114] used and the performance evaluation measures used to analyze the results.

5.5.1 Classification Algorithms

For classification task, four varied classification algorithms, i.e., SVM, DT, NB, and Logistic Regression are applied provided in ODM [109]. As discussed earlier that it is used for data mining tasks in a number of existing research works[110-112]. ODM uses Maximum Description Length (MDL) algorithm to measure significance of each
attribute. ODM provides a negative score to depict the negative role of a feature for feature reduction. ODM shows positive values for the all the proposed features.

### 5.5.2 Dataset

The choice of proper dataset is significant as it should cover diverse topics from factual to opinionative and a large number of users from all over the world share their views. Dataset of BBC Forum [114], a public discussion forum, provides positive and negative emotions for each post. It contains discussions regarding news, social issues, political and religious views for the period from July 2005 to June 2009. The statistics of the dataset are displayed in Table 33 as follows:

<table>
<thead>
<tr>
<th>Table 33: BBC Forum Dataset Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threads</td>
</tr>
<tr>
<td>Posts</td>
</tr>
<tr>
<td>Users</td>
</tr>
<tr>
<td>Avg. Posts in Threads</td>
</tr>
<tr>
<td>Avg. Users in Threads</td>
</tr>
<tr>
<td>Avg. Thread Life</td>
</tr>
<tr>
<td>Avg. Thread Length</td>
</tr>
</tbody>
</table>

Oracle data miner uses k-fold cross validation techniques, k is set to 10. For a post content, positive emotion value is from 1(low) to 5(high) and negative emotion value is from -1 to -5. The higher the value, the stronger is the emotion. The low emotion value is 1 and 2 while strong emotion value is 3, 4 or 5. A small number of posts have 5 as emotion value. To evaluate the high positive and negative emotion, a post is labelled as a Mixed-opinion post if it has both positive and negative emotion value greater than 2 (i.e., 3,4 or 5). There are total 1, 92,158 posts, out of those only 92, 159 posts are Mixed-opinion posts. A thread is labelled Mixed-opinion Thread which has the accumulative scores of the positive emotions and the accumulative score of the negative emotions greater than a certain threshold value, set as 10 in this case. There are 97,946 threads out of which 39,624 threads are Mixed-opinion threads.

### 5.5.3 Performance Evaluation Measures

The standard performance evaluation measures of Precision, Recall, F-Measure and Accuracy have been used which are calculated using the equations 10, 11, 12, 13 as
already given in chapter 4. In addition, performance evaluation measures used to evaluate classification such as Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), Lift and Cost have also been used for evaluation. The measure as briefly described as follows:

5.5.4.1 ROC

Receiver Operating Characteristic (ROC) is a metric to compare actual and predicted values in a classification model. It is applied for the analysis of binary classification to obtain in-depth understanding of the prediction capability of a classification model. ROC curve is plotted on X-Y axis where X axis denotes the false positive rate and Y axis denotes the true positive rate correspondingly. Following the ROC curve, optimal location is the top-left corner which indicates a low false positive rate and a high true positive rate [127].

ROC graph is defined by a parametric definition

\[ x = FPrate(t), \quad y = TPrate(t). \]  \hspace{1cm} (16)

Where \( t \) represents the probability threshold value, which by default is 0.5.

5.5.4.2 AUC

For comparison, ROC performance can be reduced to a single scalar value by calculating “Area Under the ROC Curve” (AUC). It measures the distinguishing capability of a binary classification algorithm. The higher its AUC value, the better it is as it shows probability that the classification algorithm ranks positive instance higher as compared to a negative instance. AUC is computed using the following formula [128].

\[ A_{ROC} = \int_0^{1/P} \frac{TP}{P} \ d \left( \frac{FP}{N} \right) = \frac{1}{PN} \int_0^N TP \ dFP \]  \hspace{1cm} (17)

5.5.6 Lift

Lift measures the level to which the predicted results of a supervised learning model are better as compared to those of randomly-generated predictions. Lift is the ratio between the percentages of correct positive classifications to that of actual positive classification in the test data. Lift is computed using the parametric definition [128]:

\[ \text{Lift} = \frac{\text{Actual Positive}}{\text{Predicted Positive}} \]
Mixed-Opinion Classification using Thread Features

\[ x = Y_{rate}(t) = \frac{TP(t)+FP(t)}{P+N}, \quad y = TP(t). \]  

(18)

5.5.7 Cost

Cost is an additional measure introduced by Oracle Data Miner. It is an indication of the damage done by an incorrect prediction and is useful for comparison of classification models. Lower cost means a high probability of confidence in the prediction ability of the classification model.

5.6 Results and Discussion

The post and thread classification results using four classification algorithms are compared using evaluation measures of Precision, Recall, Accuracy and F-measure. In addition, performance measures of ROC, AUC, Lift and Cost are used for in-depth analysis.

5.6.1 Post Results

The table 34 shows that both conventional sentiment features and proposed dialog features contribute to find mixed-opinion posts and their combined feature set shows optimal results as evidenced from the Recall results. Considering the performance evaluation measures, higher Recall and F-measure values are observed as compared to Accuracy and Precision.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment</td>
<td>NB</td>
<td>0.667</td>
<td>0.635</td>
<td>0.712</td>
<td>0.671</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.685</td>
<td>0.649</td>
<td>0.743</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.670</td>
<td>0.632</td>
<td>0.742</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.650</td>
<td><strong>0.691</strong></td>
<td>0.487</td>
<td>0.571</td>
</tr>
<tr>
<td>Dialog</td>
<td>NB</td>
<td>0.642</td>
<td>0.636</td>
<td>0.590</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.642</td>
<td>0.636</td>
<td>0.590</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.513</td>
<td>0.489</td>
<td>0.400</td>
<td>0.440</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.642</td>
<td><strong>0.636</strong></td>
<td><strong>0.590</strong></td>
<td><strong>0.612</strong></td>
</tr>
<tr>
<td>All (Sentiment + Dialog)</td>
<td>NB</td>
<td>0.682</td>
<td>0.649</td>
<td>0.731</td>
<td>0.687</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.692</td>
<td>0.636</td>
<td><strong>0.833</strong></td>
<td><strong>0.721</strong></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.602</td>
<td>0.555</td>
<td><strong>0.836</strong></td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>0.679</td>
<td><strong>0.683</strong></td>
<td>0.609</td>
<td>0.644</td>
</tr>
</tbody>
</table>

Table 34: Post Results for Mixed-Opinion Classification

“Social Web Mining based on Link and Content Analysis” 77
5.6.2 Thread Results

The results in table 35 reveal that both sentiment and dialog features find mixed-opinion threads and the proposed thread-structure features outperform both categories and the observation is consistent for all the classification algorithms.

The feature set combination provides improved results as given in Table 36. The inclusion of the proposed thread-structure features outperforms the combined feature set of sentiment and dialog. The dialog and thread-structure combination show optimal performance as compared to combination of sentiment features either with dialog or thread-structure features validating our hypothesis that proposed dialog and structural features significantly help to identify mixed-opinion threads.

Table 35: Thread Results for Mixed-Opinion Classification

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiment</strong></td>
<td><strong>NB</strong></td>
<td>0.605</td>
<td>0.616</td>
<td>0.554</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td><strong>0.705</strong></td>
<td><strong>0.718</strong></td>
<td><strong>0.672</strong></td>
<td><strong>0.694</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.611</td>
<td>0.625</td>
<td>0.549</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.591</td>
<td>0.654</td>
<td>0.384</td>
<td>0.484</td>
</tr>
<tr>
<td><strong>Dialog</strong></td>
<td><strong>NB</strong></td>
<td>0.626</td>
<td>0.657</td>
<td>0.523</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td><strong>0.659</strong></td>
<td><strong>0.640</strong></td>
<td><strong>0.720</strong></td>
<td><strong>0.678</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.617</td>
<td>0.648</td>
<td>0.509</td>
<td>0.570</td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.623</td>
<td><strong>0.735</strong></td>
<td>0.381</td>
<td>0.502</td>
</tr>
<tr>
<td><strong>Thread</strong></td>
<td><strong>NB</strong></td>
<td>0.588</td>
<td>0.613</td>
<td>0.473</td>
<td>0.534</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td><strong>0.712</strong></td>
<td><strong>0.721</strong></td>
<td><strong>0.688</strong></td>
<td><strong>0.704</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.688</td>
<td>0.689</td>
<td>0.683</td>
<td>0.686</td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.668</td>
<td><strong>0.789</strong></td>
<td>0.456</td>
<td>0.578</td>
</tr>
</tbody>
</table>

Table 36: Results to Find Mixed-opinion Threads using Feature Set Combination

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiment + Dialog</strong></td>
<td><strong>NB</strong></td>
<td>0.618</td>
<td>0.633</td>
<td>0.558</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td>0.660</td>
<td>0.642</td>
<td><strong>0.719</strong></td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.641</td>
<td>0.669</td>
<td>0.557</td>
<td>0.608</td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.641</td>
<td>0.721</td>
<td>0.458</td>
<td>0.560</td>
</tr>
<tr>
<td><strong>Sentiment + Structural</strong></td>
<td><strong>NB</strong></td>
<td>0.594</td>
<td>0.613</td>
<td>0.504</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td>0.644</td>
<td>0.634</td>
<td>0.675</td>
<td>0.654</td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.668</td>
<td>0.703</td>
<td>0.578</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.679</td>
<td><strong>0.795</strong></td>
<td>0.480</td>
<td>0.598</td>
</tr>
<tr>
<td><strong>Dialog + Structural</strong></td>
<td><strong>NB</strong></td>
<td>0.596</td>
<td>0.620</td>
<td>0.492</td>
<td>0.549</td>
</tr>
<tr>
<td></td>
<td><strong>DT</strong></td>
<td><strong>0.706</strong></td>
<td><strong>0.717</strong></td>
<td><strong>0.709</strong></td>
<td><strong>0.696</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SVM</strong></td>
<td>0.667</td>
<td>0.653</td>
<td>0.680</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LogReg</strong></td>
<td>0.682</td>
<td><strong>0.807</strong></td>
<td>0.476</td>
<td>0.598</td>
</tr>
</tbody>
</table>
5.6.3 Comparative Performance of Classification Algorithms

For post classification, decision tree outperform other classification algorithms showing the best overall results. Logistic regression presents improved outcome using sentiment features. Similar results are observed using thread-structure features also in which decision tree presents enhanced results comparing accuracy, recall and F-measure based results with those of other classification methods. Logistics regression show optimal results in precision. Naïve Bayes and SVM show relatively similar results.

5.6.4 Relative Performance of Features

To focus on feature set analysis, average of the classification results is presented in Table 37 that verify that the thread-structure features show optimal results than those of sentiment and dialog features. In addition, the combined structural and dialog feature set display improved results as compared to any combination of sentiment features. All the optimal results mentioned in bold belong to the use of thread-structure either separately or in combination with one another.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Accu</th>
<th>Prec</th>
<th>Rec</th>
<th>F-Meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment</td>
<td>0.628</td>
<td>0.653</td>
<td>0.540</td>
<td>0.586</td>
</tr>
<tr>
<td>Dialog</td>
<td>0.631</td>
<td>0.670</td>
<td>0.533</td>
<td>0.583</td>
</tr>
<tr>
<td>Structural</td>
<td>0.664</td>
<td>0.703</td>
<td>0.575</td>
<td>0.626</td>
</tr>
<tr>
<td>Sentiment+Dialog</td>
<td>0.640</td>
<td>0.666</td>
<td>0.573</td>
<td>0.610</td>
</tr>
<tr>
<td>Sentiment+Structural</td>
<td>0.646</td>
<td>0.687</td>
<td>0.559</td>
<td>0.610</td>
</tr>
<tr>
<td>Dialog+Structural</td>
<td>0.663</td>
<td>0.699</td>
<td>0.588</td>
<td>0.631</td>
</tr>
<tr>
<td>All</td>
<td>0.653</td>
<td>0.689</td>
<td>0.574</td>
<td>0.619</td>
</tr>
</tbody>
</table>

Let us compare the features using evaluation measures of Accuracy, Recall, Precision, and F-Measure. We consider the optimal results achieved regardless of the classification algorithms applied and provide the results of each feature category and combined in Figure 11 and Figure 12 respectively. Thread-structure features show the best results using all the performance evaluation measures. Dialog and sentiment feature set display comparable results. The combined feature set of Sentiment and Dialog show enhanced results, and the combined feature sets of sentiment and thread-structure provide optimal
results. The proposed thread-structure and dialog features show the best performance than all the three feature sets combined together.

![Figure 11: Classification Results using each Feature Sets](image)

![Figure 12: Classification Results using Feature Set Combinations](image)

### 5.6.5 Result analysis using ROC, AUC, Lift and Cost

The results of all the features using performance evaluation measures of ROC, AUC, Lift and Cost in Table 38 depict the importance of proposed thread and dialog features. ROC values for thread are the highest and enjoy optimal values in feature set combinations. Similar results are observed in the case of AUC and lift also. The lower cost shows the better model and the thread features show optimal results.
Table 38: Comparative analysis using ROC Accuracy, AUC, Lift and Cost Results

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>ROC</th>
<th>AUC</th>
<th>Lift</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment</td>
<td>0.610</td>
<td>0.6550</td>
<td>1.534</td>
<td>9304</td>
</tr>
<tr>
<td>Dialog</td>
<td>0.660</td>
<td>0.715</td>
<td>1.610</td>
<td>8158</td>
</tr>
<tr>
<td>Structural</td>
<td>0.713</td>
<td>0.786</td>
<td>1.747</td>
<td>6876</td>
</tr>
<tr>
<td>Sentiment+Dialog</td>
<td>0.705</td>
<td>0.772</td>
<td>1.708</td>
<td>7044</td>
</tr>
<tr>
<td>Sentiment+Structural</td>
<td>0.706</td>
<td>0.781</td>
<td>1.770</td>
<td>7042</td>
</tr>
<tr>
<td>Dialog+Structural</td>
<td>0.705</td>
<td>0.781</td>
<td>1.766</td>
<td>7062</td>
</tr>
</tbody>
</table>

5.6.6 Most Useful Features

We applied Wald chi-square statistic to evaluate the statistical significance of each feature in the classification model. The feature-wise ranking for the post and thread is presented in Table 39 and Table 40 respectively. The top-ranked post length reveals that users write lengthy content in mixed-sentiment discussions. The feature of mixed-opinions words (pMixWordsScore) is helpful to find mixed-opinion posts. The sentiment score is significant as compared to the sentiment words feature, which is understandable as the score is the sentiment valence while the words count the number of sentimental words. In dialog feature set, the top-ranked capital content verifies that it shows strong hate or negative emotion which is a common feature in controversial topic discussions. The feature of mentioning the user’s name is also significant. Quoted text is a copied content in a reply post in a question-answer thread or a user quote above text to add argument in favor of the earlier post. In other words, the direct mentioning the user is significant over quoting the text of the earlier post. The URL provision feature has low rank as a reference is given in an informative post to provide links to a detail of a fact or in reply to question and is not proven as a significant characteristic of an opinionative post.

Table 39: Ranking of Top Post Features

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Dialog</th>
</tr>
</thead>
<tbody>
<tr>
<td>numCharPost</td>
<td>boolCapital</td>
</tr>
<tr>
<td>pMixWordsScore</td>
<td>boolUsername</td>
</tr>
<tr>
<td>pSentiWords</td>
<td>boolQuotedText</td>
</tr>
<tr>
<td>boolURL</td>
<td></td>
</tr>
</tbody>
</table>
The significance of sentiment score over sentiment word is a consistent observation. In dialog features, the capital content is a significant feature. The observation in URL feature is different which a user shares to emphasize their point of view. URL existence and count are considered at posts and thread level respectively as post length is smaller. In structural features, the number of thread posts and thread-length are ranked higher and our observation is consistent with earlier work [94], which ranks the thread length and the number of posts among top features for identification of subjective threads. The proposed feature numCycleUsers which is proposed to identify conversation has 2\textsuperscript{nd} rank. The user’s consecutive posts feature is not ranked in top positions. The possible reason is that this behavior may be found in objective threads where users may share consecutive posts to explain their earlier posts, especially asking a question and explaining answer provided in the earlier post. Thread life time is ranked low as the social web can generate content any time, thus the duration may lengthy or the emotional discussion may be done in relatively short span of time.

The significance of sentiment score over sentiment word is a consistent observation. In dialog features, the capital content is a significant feature. The observation in URL feature is different which a user shares to emphasize their point of view. URL existence and count are considered at posts and thread level respectively as post length is smaller. In structural features, the number of thread posts and thread-length are ranked higher and our observation is consistent with earlier work [94], which ranks the thread length and the number of posts among top features for identification of subjective threads. The proposed feature numCycleUsers which is proposed to identify conversation has 2\textsuperscript{nd} rank. The user’s consecutive posts feature is not ranked in top positions. The possible reason is that this behavior may be found in objective threads where users may share consecutive posts to explain their earlier posts, especially asking a question and explaining answer provided in the earlier post. Thread life time is ranked low as the social web can generate content any time, thus the duration may lengthy or the emotional discussion may be done in relatively short span of time.

![Table 40: Ranking of the Top Thread features](image)

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Dialog</th>
<th>Structural</th>
</tr>
</thead>
<tbody>
<tr>
<td>tSentiScore</td>
<td>numCapital</td>
<td>numPosts</td>
</tr>
<tr>
<td>tMixWordsScore</td>
<td>numURL</td>
<td>numCycleUsers</td>
</tr>
<tr>
<td>tSentiWordsScore</td>
<td>numQuotedText</td>
<td>numThreadWords</td>
</tr>
<tr>
<td></td>
<td>numUsername</td>
<td>numConsPosts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>numThreadDuration</td>
</tr>
</tbody>
</table>

### 5.7 Chapter Summary

In this chapter, we focus to identify the mixed-opinion contents in the posts and threads of the social web forums. We propose three types of the feature sets: content based lexical sentiment feature set, content based dialog act feature set and link-based thread structure feature sets. The proposed dialog and thread specific features need no lexicons for their computations. The contribution lies in the introduction of non-lexical feature set. It is important that the lexicon features itself are novel as these are the straightforward considering the lexicon score but we introduce the word based lexical features as well as the mixed-opinion words which has not been used in the relevant literature. A real life large dataset has been used for this binary classification research problem. The obtained results using the four diverse supervised learning algorithms confirm that non-lexical
features outperform the conventional lexical features. We find out the important features which play more vital role in identification of mixed-opinion posts and discussions.
Chapter 6: Summary, Conclusion and Future Directions
Chapter 6: Summary, Conclusions, and Future Directions

6.1 Summary

Online forums are widely used sources of information and views because of the facility for its users to discuss various topics. The discussion covers technical topics which are factual and informative. On the other hands, forums contain conversation about certain opinionative topics as well. In earlier approach, dictionaries and lexicons find the subjective content. We propose various content based features which aim to identify a dialog among users positing that subjective topics have higher chance of conversation. The results validate our assumption and the proposed features outperform the lexicon methods. The proposed thread-structure based features are helpful to classify the threads into subjective and non-subjective discussions. The important features are identified and their discussion help us to analysis how each feature plays its role in the classification.

The research proposed a metric to identify influential bloggers in a blogging community using a varied feature set. The proposed metric, MIIB, consists of three modules of the productivity, the productivity, and BlogRank. Popularity refers to a blogger’s impact on other community members and the productivity compute a blogger’s activity. The BlogRank considers the significance of blogsites. The evaluation consists of feature-wise examination, module-level comparative analysis. The result evaluation of MIIB metric against the existing methods using performance evaluation measures on a dataset of real-world blog approves that the proposed metric is capable to identify the top influential bloggers.

In this work, sentiment, dialog act and thread-structure features are explored for mixed-opinion classification of posts and threads in online forums. Mixed-opinion content contains high positive and high negative sentiments and depicts the high diversity in emotional valence. The classification algorithms applied on world famous web forum confirm that the thread-structure characteristics play even more role as compared to
content-based sentiment and dialog features. The proposal of thread-structure feature set to classify threads of web forum is the main contribution in this work. The feature set not only outperforms baseline feature sets, but also decrease the overall complexity of the classification model as compared to the existing sentiment features based models without loss of performance. The proposed model may be applied to find the products having positive and negative feedback, the social issues about which people have a strong difference of views. This is also helpful to identify controversial topics or to find certain issues or policies about which the public has contradictory or bifurcated opinions. The features which play significant role in the classification are ranked and discussed.

6.2 Conclusions

For subjectivity classification of posts in online forum, we introduce the dialog features to identify the conversation among users and verify that the proposed features outperform conventional lexicon based sentiment features. The dialog and thread-specific feature sets are helpful to classify threads and discussions in online forum. The thread specific and dialog based features is computable irrespective of the language of the content in the online discussions. If a user seeks to find factual information and is not interested in public opinions and views thus subjective classification is helpful.

The modular approach of the proposed model identifies influential bloggers in the Blogosphere. The three modules play their respective roles in achieving the required role. As no gold standard exist for influential bloggers [17], thus we introduced two new approaches for evaluation. First, feature-wise ranking helps us to evaluate the over metric. Then, module-wise comparison and module role in the metric show their importance using performance evaluation measures. The used of performance evaluation measure is also a unique way. Finding influential bloggers has several potential applications such as such users can help in dissemination of the information to a large number of people.

A supervised-learning based method is applied to identify mixed-opinion posts as well as threads in online forum. The proposed dialog features successfully classifies the mixed-opinion threads as compared to lexicon based sentiment features and dialog and thread-specific features outperforms both sentiment and dialog features. The proposed features
decrease the complexity of the classification model in comparison to sentiment features based models and do no compromise the overall performance and accuracy. The proposed approach has various potential applications such as it may help us to identify such services and products about which people express positive as well as negative feedback. Also, it may help to discuss such issues about which people have higher difference of opinions.

6.3 Future Directions

Let us discuss the potential future work.

For subjectivity classification and opinion mixture identification, a potential future work is to perform topic-sensitive classification. It would be interesting to evaluate and compare the role of proposed features for diverse topics. Similarly, new content-based features may also be introduced and their comparison can be accomplished with the proposed features. In addition, dataset of various natures such as web blog can also be interesting to be analyzed with the proposed dataset. Recent works aim to find opinion in multilingual content [52, 58], it would be nice to use these features for multilingual content.

The proposed model can further be analyzed using different datasets, especially there is possibility to introduce more features and thus propose diverse modules. The new weights and their impact can also be analyzed. We intend to explore the role of sentiments in finding top bloggers. It will help us to analyze whether the top bloggers post subjective and positive content or not. A recent work takes social web values as weights for using PageRank to rank top radically influential users [77], such concept can help to use new features to find weights to find top users in blogosphere.

For subjectivity analysis and mixed-opinion classification in online forums, the possible future work can be to analysis non-lexical and thread-specific features for forum dataset. Forum users can add content in various languages and for a person; it is difficult to evaluate multilingual content. Thus, such features which are not specific to any language are helpful to analyze grammatically incorrect and instructures content found in the social web forums.
References

1. o'Reilly, T., *What is web 2.0* 2009: " O'Reilly Media, Inc.".
References

References

References

85. Turney, P.D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. in Proceedings of the 40th annual meeting on association for computational linguistics. 2002. pp.417-422.
References


102. Mohammad, S.M. and Yang, T.W., Tracking sentiment in mail: how genders differ on emotional axes. in Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011). 2011. pp.70-79


