Universiteit Leiden
Opleiding Informatica

A Topic-based Sentiment Analysis Approach to Conversation Recommendation

Name: Richard Enyinnaya
Date: 11-12-2015
1st supervisor: Frank Takes
2nd supervisor: Marvin Meeng

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Social networks have exploded as a platform of public discourse and have played a fundamental role in crisis communication, establishing human networks, mobilization of people, learning about new products and promotion of new products. The massive use of social networks by hundreds of millions worldwide has enabled the rapid propagation of information on a global scale, increasing the spread of novel information and providing a platform for everyone to share their opinion.

Interactions on social networks have become important in improving business performance, building recommendation systems, efficient customer support services, determining brand awareness, evaluating campaigns, identifying top-performing contents and anticipating events.

The research explores supervised learning, sentiment analysis and Social Network Analysis (SNA) for conversation modeling and proposes a technique based on the mechanics of interaction on a social network which takes into account user aggregated interaction sentiment polarities over time and the fraction of sentiment distributed by users for link prediction. We distinguish from previous research by considering topic-based sentiment for link prediction, by utilizing the mention graph. This approach can be utilized to create richer user social network profiles, sustainable interactions or collaborations, information propagation, development of conversation models, solving the problem of link prediction and the optimization of interactions across social networks. The approach is validated with a real-world Twitter data-set.

Acknowledgements

I would like to thank several people for making this research a success. First and foremost I want to thank my academic supervisors Frank Takes and Marvin Meeng for their feedbacks, time and effort going through many revisions of this work and their constant guidance steering my focus throughout this research. Also I like to thank the Peace Informatics Lab, Thomas Baar and Ulrich Mans for their contributions. Last but not the least, I would like to express my gratitude to my family for their non-exhaustive continued support.

Contents

Chapter 1

Introduction

In this chapter, the introductory concepts of the research will be described. Thereafter, the goal of the research, underlying motivation and the overview of research already done in sentiment analysis and information propagation is discussed. In addition, the structure and scope of the research is also described.

The effect of social networks engulfing our everyday lives has attracted a lot of interest in the research community and industry [16, 26, 56]. Social networks have played a prominent role in socio-political events by helping to speed up the process of mobilization and organization of revolutionaries, message transmission to the world and galvanization of international support [29]. Examples of cases where social networks have played a crucial role are the Egyptian revolution, the Occupy Wall Street movement, Tunisian revolution, Euromaidan (revolution in Ukraine) and the Iran revolution[1]. Furthermore, as a result of the fast development of internet technologies, mobile technology and inexpensive data storage has fostered the vast growth of users which generates an enormous amount of data each day. Therefore, organizations are now taking advantage of social networks to answer critical business questions. In addition, from the need to understand how highly complex connected systems and societies operate, the concept of networks emerged, which is a set of things or objects in which some pairs are connected by links (relationships). Social media refers to applications or websites that enable users to create or share information and ideas in virtual communities and networks. Online social networks refer to the collection of social ties with the use of internet-based social media to make connections with friends, family, classmates and customers. There are two main theories on the principles of networks; graph theory and game theory [15]. Graph theory is the study of the underlying network structure. Game theory provides models of individual behavior in settings where outcomes depend on the behavior of others, which inspired the use of information cascades [15]. Twitter, as an example, which is used in this research, is a massive microblogging social network platform with a message limit of 140 characters, sentiments, opinions, thoughts and other valuable information. As of 2014, Twitter had over 500 million tweets sent per day, and over 1 billion total users[2].

[1] http://en.wikipedia.org/wiki/Twitter_Revolution

[2] Twitter Inc.: https://about.twitter.com/company

In today's advertising and marketing sector it is sometimes said that "everything is about conversation and not broadcasting" [25]. Therefore, corporate and non-profit organizations now frequently interact with customers via so-called social network sites (SNS).

A social network is a social structure made of nodes (individuals or organizations) that are related to each other by various interdependencies: friendship, following (linking to another user and receiving the linked user's tweets) and interactions. A graphical representation of a network consists of nodes and edges (relationships) that are either directed or undirected. In a directed graph, e.g., for Twitter conversations, the set of vertices or nodes are connected together and the edges are directed from one vertex to another. An undirected graph (e.g. Facebook friendships) is the set of vertices or nodes which are connected together and whose edges are bidirectional. A mention graph is a graph where vertices are users and (directed) edges represent any interaction between the users [11].

1.1 Goals

Individuals, corporations and non-profit organizations utilize social networks to disseminate information, communicate with targeted customers or to find an audience to trigger conversations and information sharing behavior. However, it has been identified that this information or message does not trigger a conversation or reach the intended audience on respective social networks [17]. On social media sites, only a small number of users actually engage in conversations; most users are passive observers [66]. There is a fundamental need to improve and investigate the mechanics of interaction and information propagation to enhance interaction and the effectiveness of social media campaigns [2,60]. Furthermore, organisations innovate constantly by developing new products and services. These organisations are interested in what the consumers' opinions are on innovation with respect to time.

The aim of this research is to improve on addressing the issue of mechanics of interaction and conversation recommendation in social networks. This research proposes a combination of social network analysis, machine learning and sentiment analysis and Interaction Sentiment Weighting (ISW) to deduce network dynamics. Furthermore, this research also provides knowledge and experience in data mining and data analysis, i.e., knowledge discovery. In addition, opinion mining of online social network is becoming a crucial criterion for organizations in order to make decisions on whether to improve a product or service. It is becoming important to take into account what others think, after introducing an innovation into the market with the use of social network sites. So, comprehensive data mining on real-world social network based on sentiment and graph analysis will be performed.

1.2 Motivation

There exists a curiosity as to why some messages go viral on social networks and others do not. This research will investigate the mechanics of topic and sentiments of users over conversation and the role played in the propagation of information and sustainability of interactions. Human communication patterns are often associated by the fact that the sentiments associated with the conversations play a role in an individual's willingness to participate or engage with a harmonious passion. The research will deduce whether the same communication pattern applies online. Furthermore, a dominant sentiment in users' conversations could be utilized in understanding the online social profiles of users, which can be employed to enhance information dissemination. A social profile of a user is the property that defines a user's activities in an online social network e.g., Twitter. A review of the unique functions users can perform on Twitter, are the following:

Tweet: This refers to posting messages with a limit of 140 characters. The messages may contain user's daily activities, news, URLs or hashtags (e.g., #Friday).

Re-tweet: This refers to forwarding a tweet from another user to the followers, which is a prevailing mechanism for information diffusion on Twitter [36].

Mention: This is used as a marker of addressability, attention grasping or to direct a tweet to one or more specific users in order to engage in a conversation.

Follow: This refers to linking to another user and receiving the linked user's tweets afterwards [36]. The user creating the link is known as the follower, while the linked user is known as the followee.

1.3 Challenges

The preprocessing of unstructured social network data for data mining is challenging but important for effective data analysis and knowledge discovery. In addition, partitioning graphs for interesting structures from large mention graphs of Twitter data is not a trivial task. There is a lot of research on high quality graph partitioning [57, 59]. Furthermore, there is also a challenge in identifying influential nodes in community clusters of large graphs and filtering spam (irrelevant messages sent to large number users for the purpose of advertising, spreading malware etc.) nodes from graph data.

Sentiment analysis is a difficult task and poses some challenges. An example of one of these challenges is the problem of sarcasm described below.

```
@user I am not happy I woke up so early this morning had greats start of
the day.
@user some people need a pat on the head but with an airplane and that
will be woww..
```

Further challenges of sentiment analysis are the issues of abbreviations, language complexity, poor spellings, poor grammar and the several number of ways different people express their opinions. In addition, highly subjective texts are challenging to identify as positive or negative for humans, and the same applies to machines.

1.4 Previous work

The immense use of social networks has attracted a lot of research interest, for example related to viral topic detention, user influence detection, sentiment analysis and information propagation modeling [26]. Interaction propagation is an important aspect of understanding how viral messages propagate, to get early warning of a crisis and how misinformation spreads. Models have been developed to analyze online social networks with the assumption that people are influenced by actions made by individuals in their network [15, 26].

Twitter moods have been attributed to the prediction of stock prices [7], which postulates that collective public emotions can be correlated to the prediction of economic indicators. Furthermore, Twitter has also been utilized in the prediction of election results [65]. The results of the analysis showed that Twitter messages rationally reflect the offline political landscape.

Much research work has been done on the social graph of the network of followers [19, 47], identifying social network influencers [4, 8] and topic-sensitive influential users [39, 70]. Trung et al. [64] analyzed the effect of sentiment in a message toward content propagation patterns on online social networks, which is based on a fuzzy propagation mathematical model. The research was based on an online application that collects tweets and applies fuzzy mathematical propagation models on tweets that have been re-tweeted within a specified threshold (e.g., 20 times). Thereafter emotional linguistic properties of the tweets are analyzed. The result showed that tweets with emotional sentiments are re-tweeted with more frequency. Dang-Xuan et al. [12] analyzed data-sets from the political blogosphere to investigate the effect of sentiments in political weblogs in the diffusion of information. The result showed that blogs with emotional sentiments attracted more users, discussions and better information propagation in social networks. Kim et al. [35] examined the effect of sentiments in information propagation in political communication and the number of responses received from the message. The result showed that the sentiment has a positive effect on the number of responses and re-tweets. Tan et al. [63] investigated the effect of wordings in a topic controlled natural experiment. The result indicated that wordings and the topic have a positive effect on message propagation and the number of re-tweets.

Honey et al. [31] studied the use of Twitter conversations and collaborations focusing on the functions of `@`-sign (`@`mentions on Twitter) within a sample of 8,500 tweets. The study found that tweets with an `@`-sign are more focused as a marker of addressability (i.e., to direct a tweet to a specific user) and conversation interaction. In contrary, tweets without `@`-signs are more self-focused and make more general announcements [25]. Gurini et al. [28] proposed a sentiment based approach to Twitter recommendation based on the so-called sentiment-volume-objectivity-function (SVO). The results of the research showed that the approach outperforms some of the state-the-art recommendation systems.

Lu et al. [40] proposed to re-rank tweets in a user's timeline by constructing a user profile, based on a user's previous tweets and measuring the relevance between a tweet and the user's interest. The results showed that the model is effective in recommending tweets to users.

1.5 Research questions

This research focuses on addressing the following questions:

1. How can conversations on topic-based sentiments in an online social network be utilized to characterize a user, a product or service?
Conversation sentiments in an online social network will be analyzed using social network analysis techniques, natural language processing and machine learning techniques. Weighting measures are applied to the data-set to discover associated properties.

2. How can engagement on online social networks be maximized?
Here we will investigate ways in which engagements in social networks can be optimized by applying the Interaction Sentiment Weighting, social network analysis techniques and machine learning algorithms.

3. Can interaction polarities provide insight to create richer online social profiles and identify possible waves of negative and positive interactions?
Interaction sentiments will be modeled and online social profiles will be defined based on the interaction sentiments. The effectiveness will be measured in identifying possible waves of negative and positive interactions.

1.6 Approach

Data mining will be performed on the mechanics of sentiment in conversations on social media, specifically on Twitter. The remainder of this thesis consists of five chapters.

Chapter 2 introduces the key concepts and fundamental evaluation metrics that are crucial for the research being conducted. First, the chapter starts with an overview of unstructured data mining and social networks, proceeding with the discussion of rich data and networks, graph analysis evaluation metrics, sentiment analysis and machine learning evaluation metrics.

Chapter 3 starts with the description of the categories of data utilized for the research experiments. Thereafter, an outline of key data preprocessing steps for the experiments is described. Furthermore, the characteristics, empirical analysis and detected communities are presented.

Chapter 4 focuses on the proposed approach, the Interaction Sentiment Weighting (ISW) approach and results of experiments. The chapter starts with an argument to support the key reasons for the Interaction Sentiment Weighting.

Chapter 5 presents the conclusions for the experiments of the research study and then gives suggestions on directions for future work.

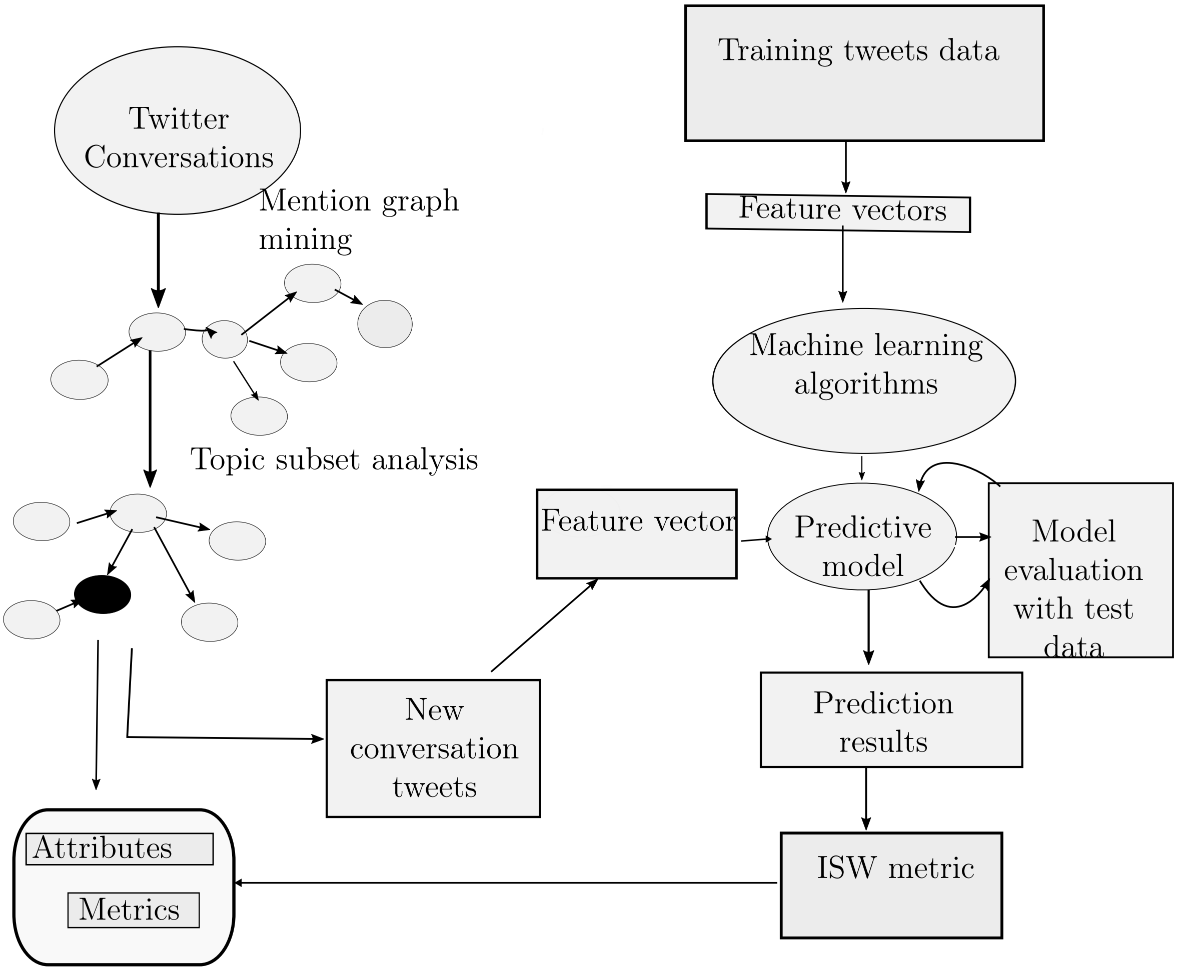


Figure 1.1: Research Approach Schematics

Figure 1.1 depicts the graphic illustration of the research approach. The mention graph is constructed and thereafter the graph is partitioned to topical subsets. The users are subjected to sentiment analysis, graph analysis and a weighting function. Furthermore, some users might be more active than others, which is denoted by the bold node. The illustrated approach helps to derive user's interaction sentiments after developing a model for sentiment classification. Thereafter, the Interaction Sentiment Weighting function is applied to the user sentiments and then an experiment is conducted to validate the effectiveness in optimizing conversation recommendation (link prediction).

Chapter 2

Theoretical Background

This chapter provides an overview of social networks, the mining of unstructured data and machine learning concepts. The chapter starts with the importance of mining the rich data of social network. The remainder of the chapter is organised as follows: Section 2.1 gives an overview of rich data and networks, Section 2.2 presents a review of social network concepts, graph mining and common evaluation metrics. Section 2.3 presents an overview of machine learning, sentiment analysis. Section 2.4 presents the common metrics utilized in evaluating machine learning algorithms in a classification task. Section 2.5 presents the representational models for text classification. Thereafter, text mining machine learning classifiers are presented in Section 2.6 and Section 2.7, Section 2.8 and 2.9 presents link prediction, Cosine Similarity and Software tools respectively.

Interactions on social networks with the combination of machine learning and social network analysis techniques can provide new knowledge in order to meet the challenges of modern society and businesses [62]. More than 1.5 billion people around the globe have an account on a social networking site, and almost one in five online hours is spent on social networks [10]. The extraction of structured data from this unstructured data as a result of the huge number of users utilizing extremely popular user-centred applications such as social blogs, social network sites and video sharing sites, all known as web 2.0 services, can provide stakeholders with useful insights to support business decisions. These applications consisting of unstructured data contain collections of valuable knowledge. For example, properties of a person, product or situations surrounding an entity or a society.

Unstructured data consists, of among others, opinions and sentiments of users which have to be processed by data mining, machine learning and natural language processing techniques before useful and meaningful information can be extracted.

2.1 Rich data and networks

Modern data-sets are enormous in size, such as social network data, which is in an unstructured format. These large data-sets arrive at high velocity and variety, and are often best represented as graphs. Data possessing these aforementioned characteristics are known as Big Data. Some of the characteristics of big data are the following.

Volume: This describes the enormity of the data to be analyzed. Modern social network data to be analyzed (e.g Facebook, Twitter, Instagram) arrives in terabytes each day and is growing at an unprecedented rate.

Velocity: This refers to the speed of data. For example, every minute 350,000 tweets are shared, Google receives 2 million search queries, users share 684,478 pieces of content on Facebook and 3,600 users share new photos on Instagram [33].

Variety: This refers to the varying types of data that are mined, e.g., social network information. This varies from traditional structured information, that fits into standard databases, where traditional queries can be run.

Veracity: This refers to the trustworthiness of the data. Big data comes in large volumes and in various forms, i.e., Twitter data consisting of abbreviations and hashtags, which often results in lack of control in data accuracy or quality.

Value: This refers to the ability to turn data into value. It is important that businesses collect and leverage large amounts of data for growth; however it is crucial to understand the business value large data will bring before embarking in big data initiatives.

Massive network data contains useful knowledge other than just vertices and edges, such as attributes of vertices, weights of edges and other properties that change overtime. Extracting rich information from large unstructured social networks data is non-trivial, but important for solving modern society and business problems.

Online social networks consist of billions of social individuals referring to others and engaging in conversations more frequently online than offline, which inspired the Internet of People (IoP) [30], which is the to personal online information collection and information modeling. This concept can be applied to personalized advertising, enhancing a company's revenue and improving the way people communicate.

Figure 2.1: Social Media and Big Data

2.2 Social networks

The rapid growth of the internet and the surprising speed and intensity of news, epidemics and financial crisis spread around the world can be attributed to the "complexity" of networks. This is based on the fact that links that connect each one of us and the way our behaviors can have an indirect ramification on the outcome of everyone else [15]. Social networking refers to collections of social ties with the use of internet-based social media to make connections with friends, family, classmates and customers.

2.2.1 Graph metrics

In this subsection, an overview of graph metrics applied for graph analysis in this research is presented. Each of the metrics quantifies properties of the graphs vertices and edges. The graph metrics provide some information about the nodes and some of the patterns that are identified and present in a graph. The following are graph metrics used in the research.

Graph: A graph consists of a set $V = V(G)$ whose elements are called vertices, points, or nodes of G and a set $E = E(G)$ of unordered pairs of distinct vertices called edges of G. Such a graph is denoted by graph $G = (V, E)$ to emphasize the two parts of a graph G.

Average degree: The average degree of a graph is the average number of outgoing and incoming edges of a node. The average degree of a *directed graph* (the edges are directed from one node to the another) is denoted as $|E|/|V|$ and for *undirected graph* (the edges are bidirectional) $2 * |E|/|V|$.

In-degree: The in-degree of a node v, where $v \in V$, written as indeg(v) is the number of edges ending at v.

Out-degree: The out-degree of a node v, where $v \in V$, written as outdeg(v) is the number of edges beginning at v.

Modularity: A metric that is designed to divide a graph into modules, communities or clusters [6]. Modularity is a highly effective approach to community detection in networks.

Community detection: A community in a network is a subset of vertices that are connected to each other. Community detection in graphs seeks to discover groups of interacting nodes within a network, and the relations between them, by using the information encoded in the graph topology [18, 72].

Average clustering coefficient: This metric measures the average of clustering coefficients (the degree to which set of nodes V in a graph tend to cluster together). The average clustering coefficient C of the whole network is the average of the clustering coefficients of all individual vertices. Average clustering coefficient is computed below, where n is the number of nodes in the network.

$$C = \frac{1}{n} \sum_{v \in G} C_v$$

Average path length: The metric computes the average distance between any two nodes in a network i.e, the number of edges one has to transverse.

Graph diameter: Measures the length of the shortest path between the most distanced nodes in a graph (i.e., the maximum distance among all pairs of nodes).

Giant component: A giant component in a network is a group of nodes that are all connected to each other, thus almost every node is reachable from almost every other.

The aforementioned graph metrics are used in this thesis. The average degree refers to the average outgoing and incoming edges of each social user (node in the mention graph). In-degree refers to users that were mentioned in a conversation (incoming edges). Out-degree refers to users mentioning others in a conversation (source nodes).

2.3 Machine learning and sentiment analysis

Machine learning has become integral to modern information technology over the past decades. With the enormous amount of data available which is (still) growing each day, there is a good reason that smart data analysis has become crucial and a necessary component for technological progress.

Machine learning is a type of Artificial Intelligence (AI) that provides computers with the ability to learn without human intervention or assistance or having to be explicitly programmed [3]. The main emphasis of machine learning is that computer programs can automatically teach themselves to grow and adapt when exposed to new data. In other words, the goal is to devise learning algorithms that do the learning automatically. Machine learning methods can automatically detect patterns in data, then use the uncovered patterns to predict future data or to perform other kinds of decision making under uncertainty. The paradigm of machine learning can be viewed as “programming by example”. The key is that it enables avoiding reinventing the wheel whenever a new problem or application is encountered. But rather than program the computer to solve the problem directly, in machine learning we search for methods by which the computer will come up with its own program based on examples that are utilized for training.

2.3.1 Sentiment analysis

Sentiment is defined as a feeling of positive or negative emotion. Sentiment Analysis is the computation and classification of subjective opinions expressed in a text, in order to determine the writer’s inclination towards a certain topic, product or service whether it is positive, negative, or neutral [67, 68].

There are different levels of sentiment analysis; the document level, sentence level and feature level. At the *document level* sentiment analysis’ major objective is to determine the overall sentiment orientation of a whole document. It assumes that each document focuses on a single object and contains opinions from a single opinion holder [14]. The *sentence level* considers each sentence as a separate unit and assumes that a sentence should contain only one opinion. A *feature level* sentiment analysis aims to create a feature, based on an opinion summary of multiple reviews. Natural language processing techniques, also called computational linguistics, is a branch of computer science that is concerned with how computational methods can aid the understanding of human language, which is utilized in sentiment analysis. In this thesis, we will focus on sentence level sentiment analysis.

2.4 Evaluation

In this section, an overview of common metrics used to evaluate machine learning algorithms and sentiment classification results are presented, concluding with the choice of metric measures that best suit the task of this thesis.

2.4.1 Cross-validation

Cross-validation is a statistical method for evaluating the performance of learning algorithms by dividing the data into two segments: one used to learn or train a model and the other used to validate or test the model [55].

In addition, cross-validation is also employed for the prevention of overfitting. Overfitting is the case whereby a model memorizes the training data by heart, which hinders it from generalizing the test data, and thus does not perform well in real-world data problems. In K-fold Cross-Validation (CV for short), the training data is split into K smaller sets and the model is trained on $K-1$ folds (within each K iterations a different part of the data is held-out for testing).

An example of a 10-fold cross-validation is shown in Figure 2.2. The split subsets of the training data are tested K times while the remaining subset serves as training data. The performance measure of a K-fold cross-validation is the average of the results in the loop.

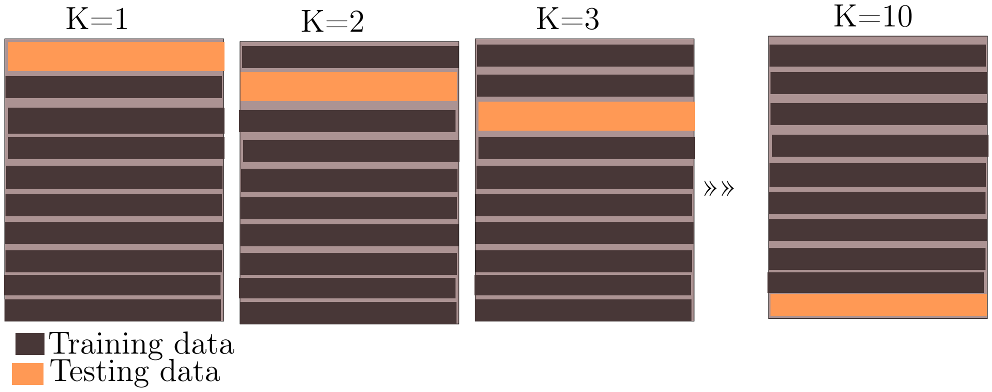


Figure 2.2: 10-fold cross-validation

2.4.2 Performance metrics

In this section, we will present performance measures that can be used to evaluate a model in a machine learning classification problem.

Performance measures (or evaluation measure) refers to a way to evaluate a solution to a problem. Performance measures are used as the criteria to evaluate learning algorithms and also used as the heuristics to construct learning models. [32]. The choice of performance measure in machine learning describes the difference between a mediocre result and a state-of-the-art usable result. Given training data consisting of input-output pairs, a model is built to predict the output from the input by fitting adjustable parameters. Once the model has been built for a classification problem, the model is evaluated on unseen data in order to determine whether it is good enough to solve the problem in the application domain under consideration.

Confusion Matrix

A confusion matrix (also called contingency table or error matrix) is a table layout that provides an easy visualization of the performance of a machine learning algorithm. An example of a confusion matrix is presented below in Table 2.3 for two possible outcomes $Positive\ (e.g., Spam)$ and $Negative\ (e.g., Not - spam)$. A binary classification problem is represented by Table 2.3 and in Table 2.4. Across the top is the observed class labels and down the side are the predicted class labels. Each of the cells (TP, FN, FP and TN) in the table contains the number of predictions made by the classifier that fall into that cell category.

	Positive (i.e., Spam)	Negative (i.e., Not-spam)
Positive (spam)	TP (True Positive)	FN (False Negatives)
Negative (Not-spam)	FP (False Positive)	TN (True Negative)

Table 2.3: confusion matrix

	Positive (i.e., Spam)	Negative (i.e., Not-spam)
Positive (spam)	Correct positives	Missed correct positives
Negative (Not-spam)	False alarms	Correct negatives

Table 2.4: Definition

Accuracy: Defined as the fraction of correctly classified instances of documents compared to the total number of instances/documents. This is the simplest and most common classification measure and computes the overall effectiveness of a classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Positive predictive value, the proportion of predicted positives which are actually positive. This measure refers to the percentage of positive examples that are correct. Furthermore, precision answers the question of: How often does a system get positive correct answers? Precision returns a value between 0 and 1, where 0 is the worst value meaning none of the recommended conversations (links) match the user and 1 means that all recommended conversations are also of the users.

$$Precision = \frac{TP}{TP + FP}$$

Recall: The proportion of actual positives that are predicted positive. This measure refers to the percentage of correct examples that are positive. Furthermore, recall answers the question of: How often does a system correctly identify positive examples when the system encounters them? Recall returns a value between 0 and 1, where 0 is the worst value and 1 is the best value.

$$Recall = \frac{TP}{TP + FN}$$

F-measure: The harmonic mean between precision and recall. This is a combined measure that assesses the precision-recall [42]. F-measure returns a value between 0 and 1, where 0 is the worst value and 1 is the best value. An F-measure is presented below:

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Matthews Correlation Coefficient: The Matthews Correlation Coefficient (MCC) is used in machine learning for the evaluation of the quality of a binary classification. The MCC returns a value between -1 and +1, where +1 represents perfect prediction, 0 random prediction and -1 an inverse prediction.

$$Matthews\,Correlation = \frac{TP.TN - TP.FN}{((TP + FP)(TP + FN)(TN + FN))^{1/2}}$$

Mean Squared Error (MSE): MSE is an evaluation metric used for the evaluation of the quality of a predictor result. Let n be the number of instances, $\widehat{y}$ be the predictions and y be the observed values inputs to the function that generated the prediction, then MSE is:

$$\frac{1}{n}\sum_{i}(\widehat{y_i} - y_i)^2$$

Root Mean Squared Error (RMSE).

$$\sqrt{\frac{1}{n}\sum_{i}(\widehat{y_i} - y_i)^2}$$

In this thesis, the evaluation metrics that will be utilized are precision, recall and F-measure. Goutte et al. [24] studied the assessment of the confidence of precision, recall and F-measure in the evaluation of Information Retrieval (IR) systems or Natural Language Processing (NLP). The research indicates that precision, recall and F-measure gives a complete view of a system's performance in the evaluation of IR systems or NLP. So we use: Precision, Recall and F-measure.

2.5 Representation models for text classification

Text classification is the automatic classification of text into categories. Text classification is a popular research topic, due to its numerous applications such as filtering spam of emails, categorizing web pages and analyzing the sentiment of social media content [22]. We consider how to represent this textual data in numeric representation to be used for machine learning classification. There are various approaches to tackling this problem. The following subsections describe the approaches.

2.5.1 Bag of words

The bag of words model learns the vocabulary from all documents, disregarding grammar and word order. It then models each word by counting the frequency of each word. The resulting feature is a row vector which contains the frequency of each word occurrence, which is then utilized for training the classifier. For example:

Sentence 1: The dog is wearing a hat.
Sentence 2: The dog is having the lunch.
Sentence 3: The cat is chasing the dog.

A vocabulary is learned and created from these sentences with unique terms {The, dog, is, wearing, a, hat, having, lunch, cat, chasing}. The unique words learned from the vocabulary of the sentences are given identifiers 1 to 10.

	The	**Dog**	**is**	**wearing**	**a**	**hat**	**having**	**lunch**	**cat**	**chasing**
Sentence 1	1	1	1	1	1	1	0	0	0	0
Sentence 2	2	1	1	0	0	0	0	1	0	0
Sentence 3	2	1	1	0	0	0	0	0	1	1

2.5.2 N-grams or bags of word sequences

Bags of word sequences are also known as the *N-gram model.* This refers to a bag of frequent word sequences. N-gram is similar to bag of words, however the main difference is that instead of computing the frequency of occurrence of each word, the frequency of word sequences is computed. Word sequences can be in the form of bigrams or trigrams, or more. For example, Let $n = 2$ (bigram).

Sentence 1: The dog is wearing a hat.
Sentence 2: The dog is having the lunch.
Sentence 3: The cat is chasing the dog.

The list of bigrams of $N = 2$ learned from the vocabulary are listed below, which is done by moving each word forward to generate the next bigram.

{The dog, dog is, is wearing, wearing a, a hat, is having, having the, the lunch, the cat, cat is, is chasing, chasing the, the dog}

Bigrams and trigrams are feature spaces for training a classifier in a supervised learning algorithm for text classification. The use of bigrams or trigrams does not necessarily yield significant improvement to the classifier [5] compared to the Bag of words.

2.5.3 Word2Vec deep learning approach

Deep learning refers to machine learning techniques or algorithms that exploit multiple non-linear layers for information processing that can learn from feature hierarchies [13]. Word2vec is a distributed representation of words in a vector space learned by neural networks, using the word2vec algorithm created by Google [46]. Word2vec does not require labelled data to create word representations and it is computationally efficient as it learns more quickly from billions of words than other neural network approaches. In the distributed representation of words (Word2vec), words with similar meanings are in the same clusters. For example a well-trained Word2vec model will produce the following " King - man + woman = queen " based on computing vector representations of words. The Word2vec architecture is outside the scope of this thesis.

2.6 Text mining classifiers

In this section common text mining classification algorithms used in machine learning will be presented. The Naive Bayes, Logistic Regression and Support Vector Machines (SVM) are described.

Text mining is a field of computer science with connections to natural language processing, machine learning, data mining, knowledge management and information retrieval, that aims to retrieve useful information from unstructured text [52]. Text mining classification is the automated analysis or the process of classifying natural language text into categories. The classification of text has great practical importance due to the large amount of new online text data that becomes available daily through the internet [48], such as electronic mail, social media, digital libraries and corporate databases. Text classification consist of the classification of documents or text into topics or sentiments. A classification task involves the separation of data into a training and test set, where each of the training instances has a "target value" also called class labels and set of "attributes" known as features which is utilized to develop a model based on the training data [38]. The developed model is used to predict the target values (class labels) of the test data based on the test data attributes (features).

The machine learning algorithms used in the experiments are Naive Bayes, Logistic Regression and Support Vector Machines. Naive Bayes algorithms are based on the use of

Bayes' theorem with the assumption of independence of feature pairs. It is one of the most effective and efficient machine learning algorithms for inductive learning despite the independence assumption between features that is often not applicable in the real-world [74].Variants of Naive Bayes such as multinomial Naive Bayes use the Naive Bayes algorithm for multinomially distributed data used in text classification.

Logistic Regression or maximum-entropy (MaxEnt) is a linear model for classification and not regression, despite the name. In a Logistic Regression model, the probabilities of possible outcomes of a single trial are modeled using a logistic function [61]. Logistic Regression (LR) is used in various applications such as document classification and natural language processing [73].

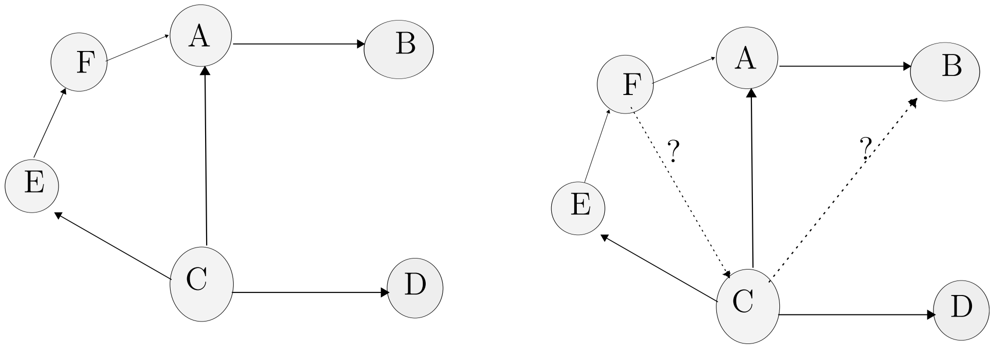
Support Vector Machines (SVMs) are widely used and popular machine learning algorithms for classification [9, 38]. In supervised learning, a training set consisting of instances of class label pairs, the Support Vector Machine (SVM) needs a solution to an optimization problem [38]. The training features are mapped to a higher dimensional space. The SVM then finds a separating hyper-plane with the maximal margin in the higher dimensional space.

2.7 Link prediction techniques

Given a snapshot of a social network, for example as in Figure 2.9 (left), can we infer which new interactions are likely to occur in the near future? This question is a link prediction problem [37]. Link prediction is an important active research area [20, 21] for analyzing social network evolution, which is relevant to a number of current application domains. Some examples of relevant applications of link prediction are; identification of spurious interactions, monitoring terrorist networks (to infer possible interactions without direct evidence), information retrieval, e-commerce, bioinformatics, and so on. In addition, link prediction is utilized in online social networks for friend suggestion [37]. Link prediction is formalized as; given a snapshot of a social network at time t, can it be predicted which new connections among its members are likely to occur (added to the network) in the future at time $t + 1$ [27, 37].

A social network is dynamic, it grows and changes with time through the addition of new edges or the deletion of existing edges. The addition of new edges indicates the appearance of a new interaction or collaboration in the underlying social network structure as in the case of the Twitter mention graph. One approach to link prediction is to make predictions of links based entirely on the structure of the network. A survey of different graph proximity measures utilized for link prediction and the comparison random predictor to graph proximity approaches is presented in [37]. Furthermore, Popescul et al. [51] proposed the use of a structured logistic model and the use of relational features from data stored in relational databases to predict links. In addition, O'Madadhain et al. [50] proposed a conditional probability model based on event-based network data over time attributes and structural features.

Link prediction for new edges is challenging. A result of interesting link data features are sparse and prior probability of a link is usually small [21]. Some of the challenges of link prediction are presented in [54].



Snapshot: Time t

Snapshot: Time t +1

Figure 2.9: Link prediction problem

To do supervised learning for link prediction, a labelled data-set has to be produced. A social network of $G = (V, E)$ in which the graph is a directed graph, the vertices (V) are users, edges (E) represent any interaction (@-mention) between the users at different points in time, with no self connections. The network is analyzed at two different time periods t and t' where $t < t'$. The features of the social network upto time t are utilized to predict new edges (links) that will be formed in the time upto t'. Let $G[t]$ be the sub-graph of G containing all the nodes and edges of time period upto t, this is referred to as the training interval data-set. Let $G[t']$ be the subgraph of G also containing all the nodes and edges of time upto t', this is referred as the test interval data-set. Clearly, $E^{train} \cup E^{test} = E$ and $E^{test} \cap E^{train} \neq \emptyset$.

2.8 Cosine similarity

Assume we model using feature vector. To measure the similarity between two users, the cosine similarity is used. The cosine similarity is a measure that calculates the cosine of the angle between two vectors. Cosine similarity is commonly used to find the similarity between two entities. A cosine similarity of 0 shows that two users are dissimilar because the angle between the users is nearly 90 degrees, as shown in Figure 2.10. A cosine similarity value that is near 1 indicates that the users are similar and the angle is pointing in the same direction. A cosine value near -1 shows that the users are in opposite directions.

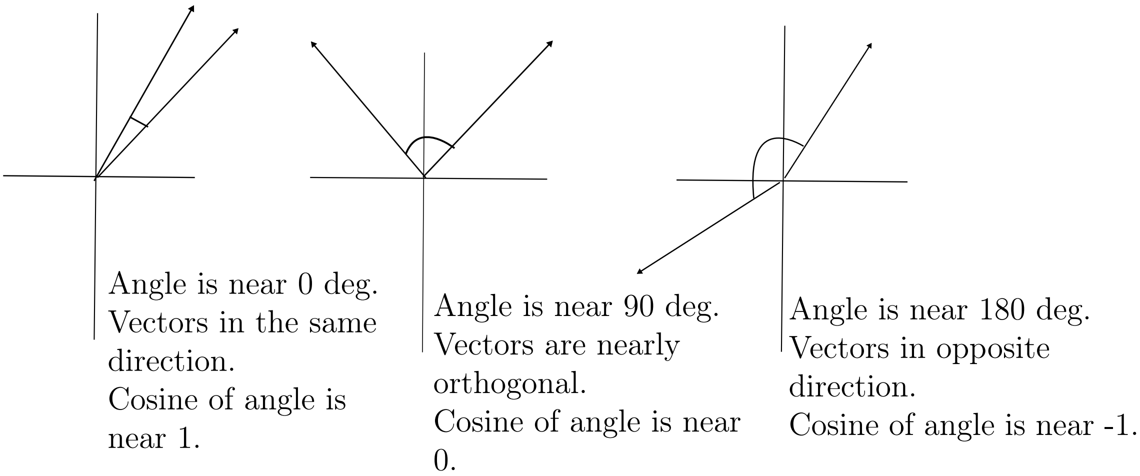

Angle is near 0 deg.
Vectors in the same direction.
Cosine of angle is near 1.
Angle is near 90 deg.
Vectors are nearly orthogonal.
Cosine of angle is near 0.
Angle is near 180 deg.
Vectors in opposite direction.
Cosine of angle is near -1.

Figure 2.10: The Cosine Similarity values for different vectors, 1 (same direction), 0 (90 deg.), -1 (opposite directions)

Given two users u_i and u_j and the profiles $I^n_{u_i}$ and $I^n_{u_j}$ (feature vectors) the cosine similarity between the two users is defined by:

$$\begin{aligned} \textit{Cosine similarity} &= \frac{I^n_{u_i} \cdot I^n_{u_j}}{||I^n_{u_i}||||I^n_{u_j}||} \\ &= \frac{\sum\limits_{n=1}^{n} I^n_{u_i} \times I^n_{u_j}}{\sqrt{\sum\limits_{n=1}^{n}(I^n_{u_i})^2} \times \sqrt{\sum\limits_{n=1}^{n}(I^n_{u_j})^2}} \end{aligned} \tag{2.1}$$

2.9 Software tools

There are a lot of software and programming tools for data mining, data processing, data analysis and text classification. The various software tools presented below are used in this thesis and have different strengths and weaknesses. The tools described in this section are all open source.

NLTK: Natural Language Toolkit is a Python library for building Python programs to work with human language data. NLTK contains tools like stemming, stop words and other useful tools for computational linguistics using Python. (`http://www.nltk.org/`)

Scikit-learn: Scikit-learn is a Python library for machine learning, containing state-of-the-art machine learning algorithms. Machine learning algorithms such as Naive Bayes and Support Vector Machines implemented in Scikit-learn were used in this research. (`http://scikit-learn.org/`).

Pandas: Pandas is a Python library that provides efficient data structures and data analysis tools for the Python programming language. Pandas dataframe and data analysis tools were employed in this thesis. (`http://pandas.pydata.org/`).

Gensim: Gensim is an open source Python library that is efficient for unsupervised semantic analysis of plain text. Word2vec in Gensim library was employed in this research. (`https://radimrehurek.com/gensim/`).

Gephi: Gephi is an interactive visualization and exploration platform for the study of networks and complex systems. Gephi was used to visualize the mention graph in this research. (`http://gephi.github.io/`).

NetworkX: NetworkX is an open source Python library for the study of networks and its underlying dynamics, functions and structure. NetworkX was employed in research to study the network structure. (`https://networkx.github.io/`).

Chapter 3

Data

In this chapter the data used for the research experiments will be presented. The data processing techniques, data characteristics, empirical analysis and the mention graph are discussed.

Three categories of data-sets are utilized in the research. The first category is the data-set utilized for the research experiments, data analysis, classification and prediction. The data consists of tweets for a period of six months from July 2009 to January 2010, which is about 20-30% of all public tweets during the particular period collected by Yang et al. [71]. The data-set contained 476 million tweets in several languages. In uncompressed format the data size is approximately 68 Gigabytes on disk. However, only English tweets were considered by selecting only tweets that contained ASCII characters resulting in 457,791,122 tweets. Each tweet has the following information: time of the tweet (T), author (U), and the tweet content (W), for example, T: 2009-06-07 02:07:42, U: http://twitter.com/caitlin and W: cleaning my room.

The second category is the data used for developing a sentiment analysis system. The data was created by Alec Go, Richa Bhayani and Lei Huang [23] from Stanford university. The data was used for training and the validation of the sentiment analysis system. In this work, two data-sets were collected for training and testing of the machine learning algorithms by querying the Twitter non-streaming API for messages that contained emoticons between April 2009 and June 2009. This was employed by Alec et al. [23] for sentiment classification. The training data contained 800,000 tweets with positive emoticons and another 800,000 tweets with negative emoticons. The validation data-set was manually annotated with class labels and consists of 177 negative tweets and 182 positive ones. Test tweets were collected by looking for messages that contained a sentiment. The data contained the following information: polarity of the tweet (positive or negative), the ID of the tweet, the date, query used, the user that tweeted and the text of the tweet.

The third category of data-set is a set of 5,389 neutral tweets of multiple brands or products from Crowdflower[1], an open data library containing a repository of data-sets collected or enhanced by CrowdFlower's contributors and made available for anyone to

[1] `http://www.crowdflower.com/data-for-everyone`

use for research and to uncover new insights. This data-set is used to enrich the sentiment analysis system as a result of the absence of neutral tweets in [23]: the original data-set used in developing a sentiment classification system.

3.1 Preprocessing

Data preprocessing refers to a data preparation technique that involves the transformation of raw data into a usable format. This section describes the various techniques undertaken in a text data preprocessing task in order to transform data into a more usable format. Each of the techniques described in the following subsections are still active fields of research. Techniques in data processing have varying strengths and weaknesses and the effectiveness is determined by performing experiments.

In order for a machine learning classifier on Twitter data to be effective, data preprocessing has to be performed to reduce the impact of noise due to the presence of irregular words, tweet's shortness and other creative language characteristics. The following subsections highlight the main data preprocessing steps performed on the data-sets utilized for the experiments.

3.1.1 Cleaning text

Data cleaning or scrubbing is one of the most important step in data processing [34]. This refers to the detection and removal of errors and inconsistencies from data in order to improve the quality of the data [53]. Data cleaning is labour intensive, expensive and often takes more than 60% of the time used for data mining [34].

Web data comes in various formats and often contains URLs, missing data (data expected is unavailable), erroneous data (incorrect spellings) and duplicated records as presented in Table 3.1. The wrong form of data preprocessing can result in damaging the information contained in the data [34].

In this research, usernames present in a web document, such as Twitter messages were removed as this does not give a unique meaning to a tweet for text data analysis and does not indicate any sentiment on a tweet. Moreover, in the mention graph usernames were not removed. In addition, a filter on non-English characters was applied. Uppercase text increases the number of unique words. The conversion of uppercase characters to lower case characters reduces the number of unique words in a document while still retaining the meaning of the words, thereby enhancing the efficiency of text classification.

User	Mention	Message	Type
Rosen	Persie	the yar is grt	erroneous
Peter	Messi	tonight is a Chelsea game	duplicate
Mary		great weather	missing data
Oliver	Aaron	fun game by Barcelona FC	correct
Peter	Messi	tonight is Chelsea game	duplicate

Table 3.1: Examples of missing, erroneous and duplicated data

3.1.2 Punctuation, numbers and stop words

The approach to text cleaning depends on the type of data problem that is being solved. For some problems it is important to remove punctuations and for others it is not. In this research, the data problem is sentiment analysis, hence it is possible that **:-)**, **(-:** or **!!!** may carry sentiment and can be treated as words. The same applies to treating numbers for sentiment analysis. Numbers can be treated as words, removed or replaced with a placeholder. In the experiment punctuations, emoticons and numbers were removed as the MaxEnt and SVM classifiers tend to put a large amount of weight on the emoticons which affects classification performance [23].

Words are the properties that describe a sentence or a document, hence are important in sentiment analysis. Sentiment classification on Twitter is often affected by the noisy nature (i.e., abbreviations and irregular word forms) of tweets [58]. Words such as "a", "or", "at", "are" and "the" often occur in every sentence and therefore they are not descriptive. Most frequent words can be identified by setting a maximum threshold on the number of times a word appears in a document. Stop words can be removed by setting a word frequency above a set maximum threshold or by checking word presence in a stop word list. In this research, stop words were removed by checking against a stop word list. Stop word lists are publicly available for most languages and an English stop word list was used in this research.

3.1.3 Stemming and Lemmatization

Stemming refers to the removal of suffixes or prefixes in order to derive the base, root or stem of a given word. Lemmatization refers to the proper use of vocabulary words and morphological analysis of words, aiming to remove the inflectional endings only and to return to the base or dictionary form of a word, which is known as lemma [42]. The major aim of stemming and lemmatization is to reduce the derivations of a word to a common base or root form. In this research, stemming was applied and the effectiveness was investigated (Section 4.4) in improving a sentiment classification system. Some of the challenges of stemming and lemmatization are accuracy, language complexity and computational time. The most used and common algorithm for stemming in English that has been repeatedly shown to be effective is Porter's Algorithm [42]. Other stemming algorithms are Lovins stemmer and Paice stemmer [42].

Word	**Stemming**
having	hav
provision	provis
owed	owe
multiply	multipli

Table 3.2 Examples of stemmers

Word	**Lemma**
having	have
churches	church
wives	wife
provision	provision

Table 3.3 Examples of Lemmatizers

3.1.4 Tokenization

Tokenization is a method of splitting text into individual words, phrases or other useful elements, known as tokens. Tokenization was applied in the research. Tokenization has some challenges, such as what are the right tokens to use, how to handle various uses of apostrophes, how to tokenize on whitespaces and whether to ignore punctuations. Example of tokenization:

Aren't we going to the capital ?

What is the correct form for aren't?

Are n't we going to the capital .

Aren t we going to the capital .

Aren ' t we going to the capital .

Tokenization is challenging and often requires the language of the document to be known (in this research, Twitter messages in English) such as tokenizing email addresses, web URLs and numeric IP addresses.

3.2 Study of data characteristics

This section presents the Twitter @-mentions data statistics utilized for the experiments conducted. The section starts with a general statistical description of the interactions in the whole data-set and proceeds to presenting the partitioned Twitter topics, then, common graph evaluations of the underlying mention graph are done.

3.3 General statistics of raw data

To effectively utilize the data, there is a need to understand some of the characteristics of the data-set. The characteristics found on the data show that 48% of all tweets (476 million) contain at least one @-mention. Table 3.2 shows the data-set statistics of the whole data containing at least one @-mention. On the order hand, a standard deviation of Twitter interactions of the data-set with a value of 0 indicates that the data points are close to the mean. The 25^{th} percentile (25%) and 50^{th} percentile (50%) and 75^{th} percentile (75%) of the whole @-mention data are 1, 3 and 12. A value 75^{th} percentile equal to 12 denotes that 75% of Twitter mentions in the whole data is below 12 @-mentions.

Total number of interactions	113,671,366
Total number of unique users	7,299,949
Mean	27.0
Standard deviation	189.0
Min	1
25%	1
50%	3
75%	12
Max	260,423

Table 3.2: Data statistics whole Twitter data of at least one @-mention

Table 3.3 shows the trending topics used for the research. The trending topics were manually selected from the data-set based on the frequency of hash-tags used by social users in interactions, collaborations or conversations on Twitter. The partitioned data is representative of the whole Twitter data-set (the use of hashtags are used to group tweets into topics[2]). This provides an efficient way to validate the research hypothesis (questions presented in Section 1.5) of Twitter interactions taking place in an online social network.

[2] https://media.twitter.com/best-practice/using-hashtags

Technology	**Political Events**	**Sports**
Google wave	IranElection	Super Bowl
Snow Leopard	G20	Lakers
Tweet deck	Obama	Cavs (Cleveland Cavaliers)
Windows 7	healthcare	Superbowl
CES	Tehran	Chelsea
Palm Pre	Mousavi	NFL
Macworld	Ahmadinejad	UFC 100
E3	Iran	Yankees
amazonfail		

Table 3.3: Trending topics

The above topics were analyzed to derive insights about the strengths and dynamics of topics and how each changes over time using Interaction Polarity Modeling (Section 4.4.1). The Interaction Polarity Modeling of topics was employed to model the wave of interactions sentiments in an OSN in order to characterize a user, a product or service. Approaches to topic sentiment analysis takes the whole summary of sentiment in weblogs ignoring the subtopics [45]. The Interaction Polarity Modeling (Equation 4.1) approach enables more in-depth analysis of sentiments and the underlying subtopics over time.

3.3.1 Mention graph statistics

In this section we describe mention graph statistics that are relevant to the research.

Some graph evaluation metrics that are relevant to the research are described in Table 3.4 and Table 3.5 below. The table show the metric values for average degree, modularity, average clustering coefficient of the partitioned topic subsets. The edges of Technology and Sports in Table 3.4 show that the data-set is sparse as a result of loosed components in the network.

	Full graph	**Technology**	**Political Events**	**Sports**
Node	14,435,577	175,792	145,397	136,473
Edges	50,029,174	140,868	170,857	115,654
Average degree	3.470	0.801	1.175	0.847
Modularity	-	0.953	0.767	0.932
Avergae clustering coeficient	-	0.003	0.004	0.003

Table 3.4: Full mention graph metrics

	Technology	**Political Events**	**Sports**
Nodes	65,528	76,077	55,761
Edges	74,480	132,039	67,521
Average degree	1.137	1.736	1.211
Modularity	0.909	0.733	0.886
Diameter	25	27	23
Average clustering coefficient	0.006	0.006	0.005
Average path length	9.64	6.735	7.916

Table 3.5: Giant component mention graph metrics

The average degree describes the average number of outgoing and incoming edges of a node. The average degree of Technology and Sports with values 0.801 and 0.847 respectively in Figure 3.4, shows that the nodes have a small number of incoming and outgoing edges compared to political events. The modularity value describes how densely connected the nodes are within modules. A Technology modularity value of 0.953 indicates a dense connection between the nodes within modules but sparse connections between nodes in different modules. On the other hand, the Table 3.5, the mention graph giant component, the group of nodes that are all connected to each other shows a high average degree.

3.3.2 Degree distribution

This section describes the distribution of the average degree based on partitioned topics described in Table 3.4.

The in-degree and out-degree (described in Section 2.2) of the data-set distribution were explored in order to get an insight into the properties of nodes. The plot of in-degree refers to the users mentioned in a conversation; while the plot of out-degree refers to users that are interacting and mentioning others in a conversation. Furthermore, plotting the degree distribution on a logarithmic scale as seen in Figure 3.1 and Figure 3.2, reveals a long tail of the degree distribution of interactions in Twitter. It shows that a lot of the nodes have a small degree and few nodes have a high degree. These "centres of activity" are nodes that are orders of magnitude larger in degree than most other nodes, a known phenomenon in power law networks known as scale free networks [69].

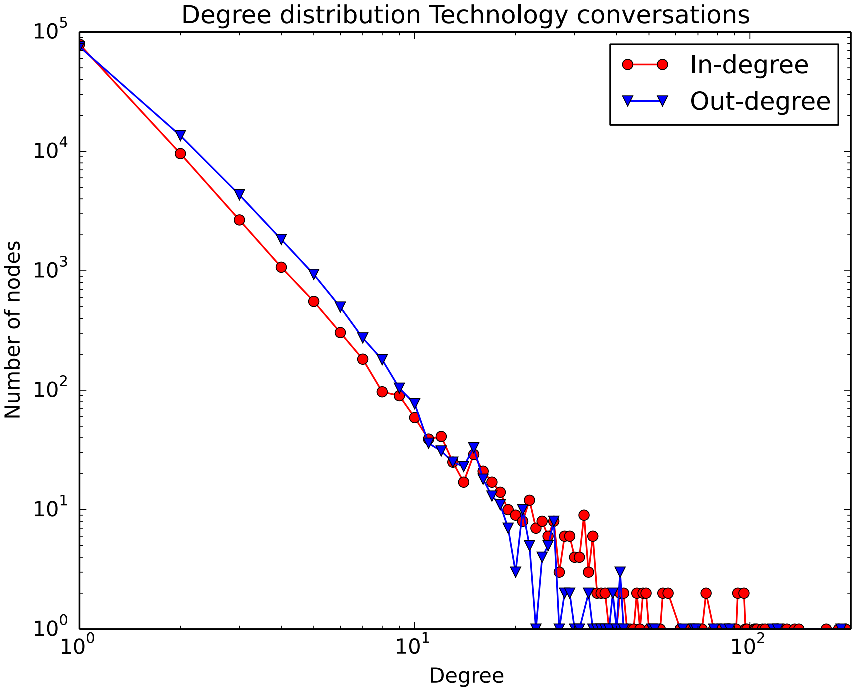


Figure 3.1: Degree distribution of technology interactions

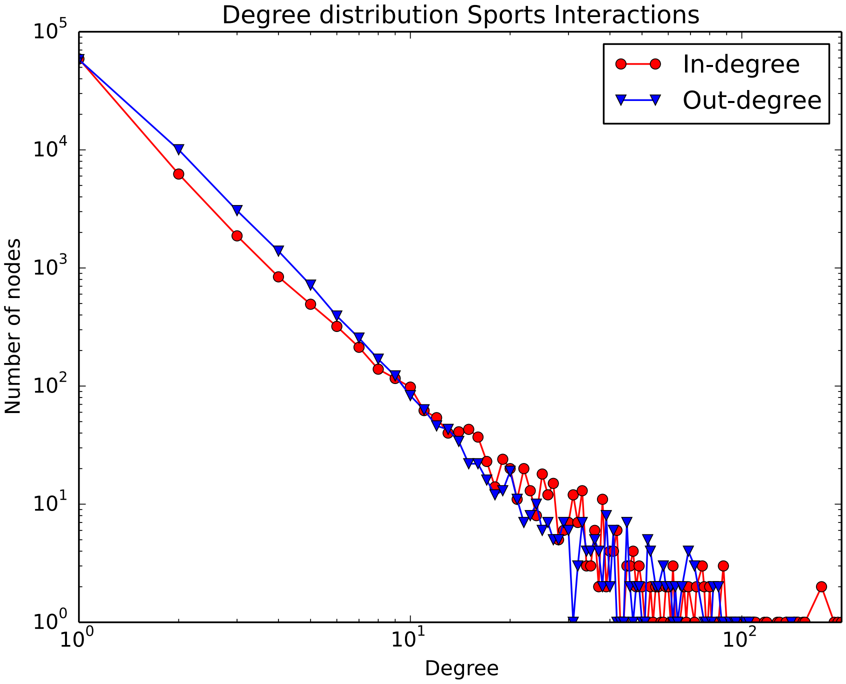


Figure 3.2: Degree distribution of sports interactions

3.4 Empirical analysis

In the following subsections, the time of interactions on social networks such as Twitter is explored. In addition, the frequency distribution of interactions is presented.

3.4.1 Twitter conversation time series analysis

The month most users engage in conversations is explored by using the whole 476 million public tweets collected by Yang et al. [71]. Figure 3.4 shows that users engage more actively in interactions in the month of August followed by September, which probably can be attributed to summer holiday period. This is measured in a span of six months, from June to January. Furthermore, the time users engage more actively in interaction is explored using time stamp based on Coordinated Universal Time (UTC), the 24-hour standard. Figure 3.3 shows the time users engage more actively in interactions is at 00.00 hours. In addition, the time most users engage least in social networks is at 9.00 hours. The data consists of tweets from all around the world with different time zones, therefore further research is needed to draw valid conclusions.

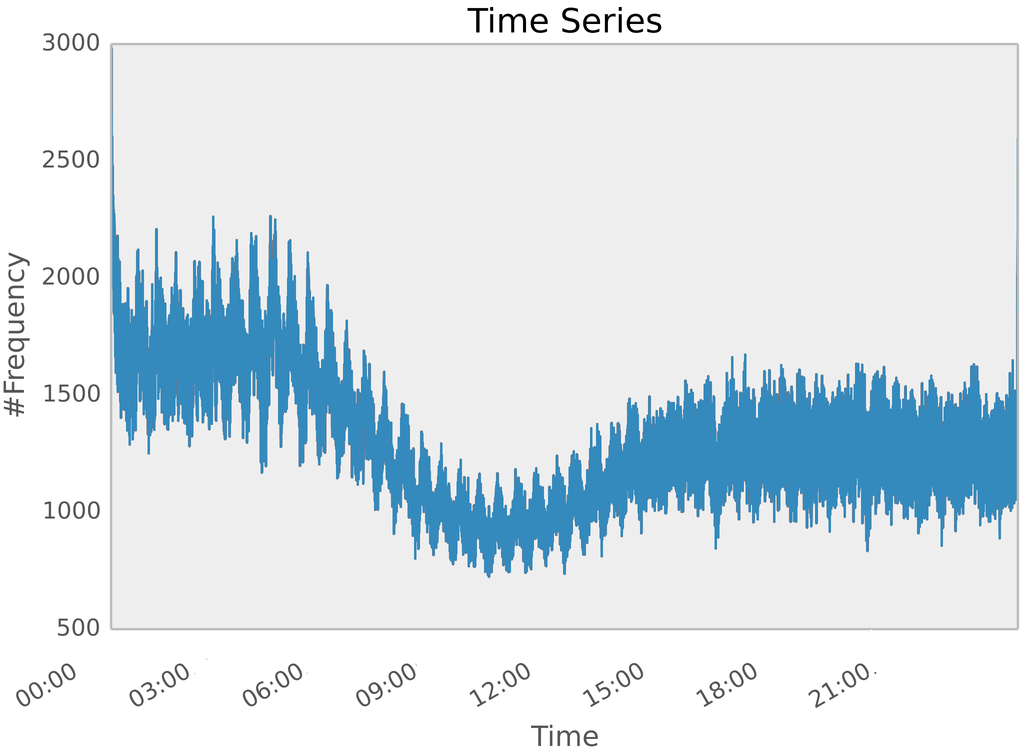


Figure 3.3: Time series interactions distribution

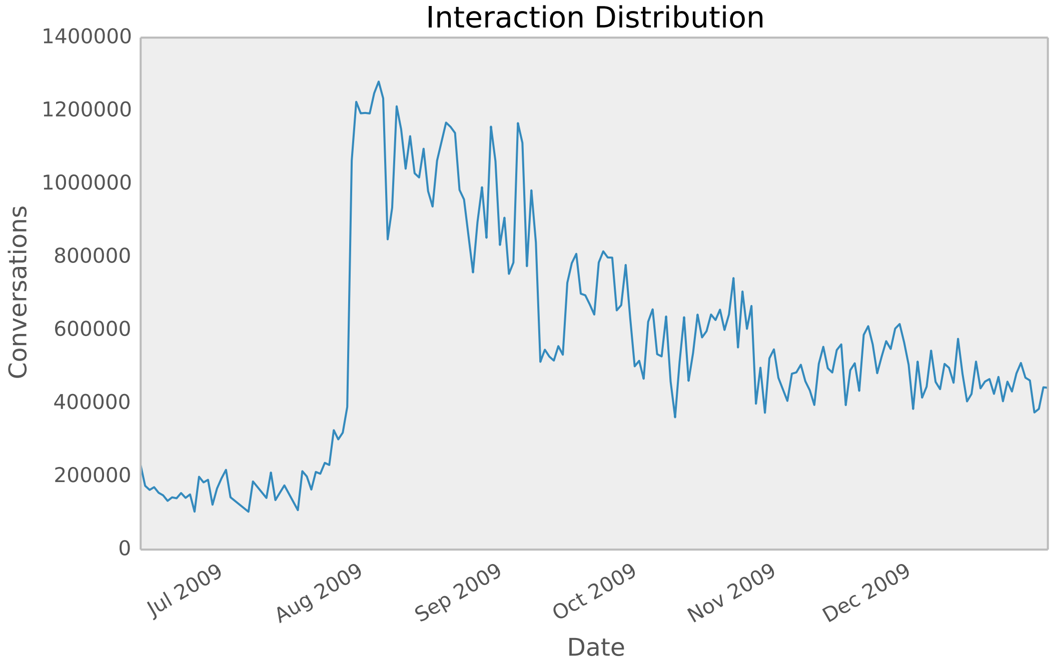


Figure 3.4: Interaction dates distribution

3.4.2 Interaction frequency analysis

The interactions distribution on Twitter is plotted in Figure 3.5 which shows the number of users mentioning others and engaging in direct conversations (sending a direct message to another social user on Twitter). This shows that Twitter is a direct conversation platform for engaging interactively other users.

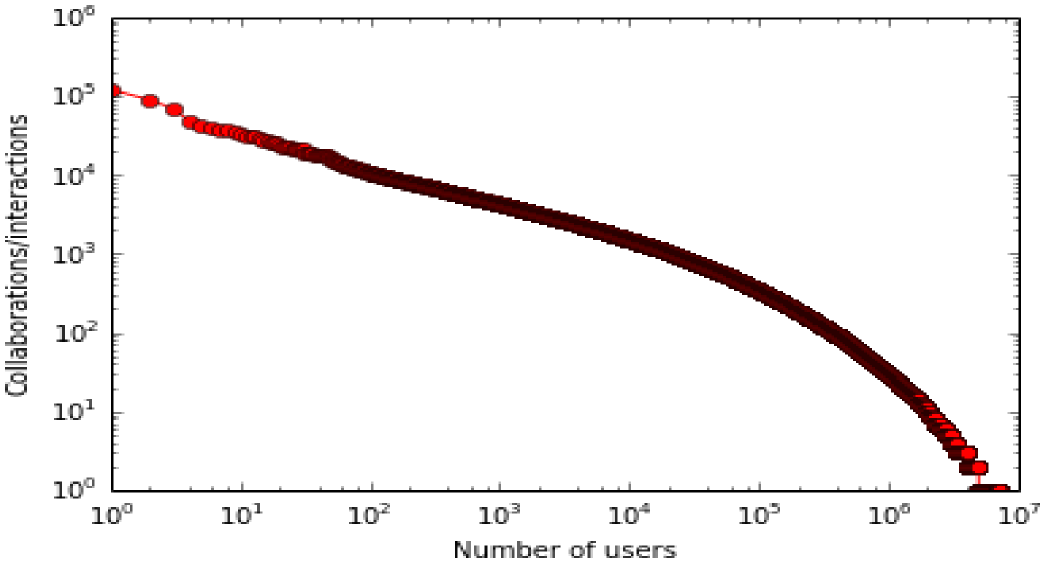


Figure 3.5: Interactions distribution

3.5 Mention graph

From the modularity algorithm in Gephi and using the Force Atlas [49] algorithm for visualization, several communities can be identified, as shown in Figure 3.6 for technology sampled data in Table 3.4. The early adopters of a new technology such as "windows" and "mac" can be identified. Modularity is used to discover communities consisting of groups that interact more with each other than with the rest of the network. Expert users of a technology and opinion leaders can be identified by the decomposition of users into clusters.

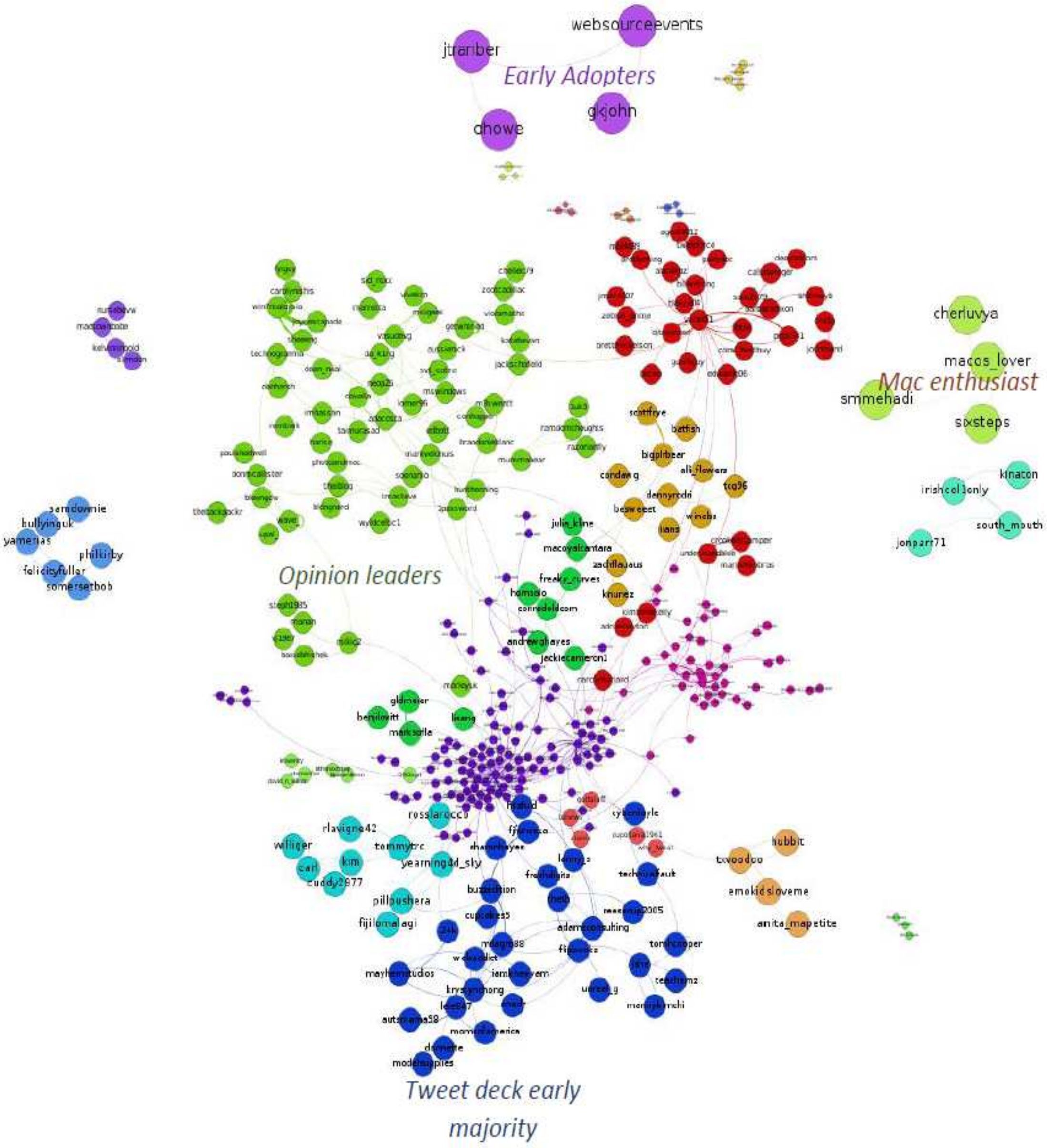


Figure 3.6: Visible communities of Technology users

3.5.1 Out-degree in mention graph

The out-degree of users is investigated to identify users who engage the most in conversations in social networks such as Twitter. This property can be used to solicit a direct answer to a question from another user, reply to a conversation and also for the collaboration. In Figure 3.7 users that engage and interact with others more frequently are shown and denoted by larger circles (degree centrality determines node size). This characteristic can be effective in identifying information seekers, opinion leaders and people who need help during a crisis.

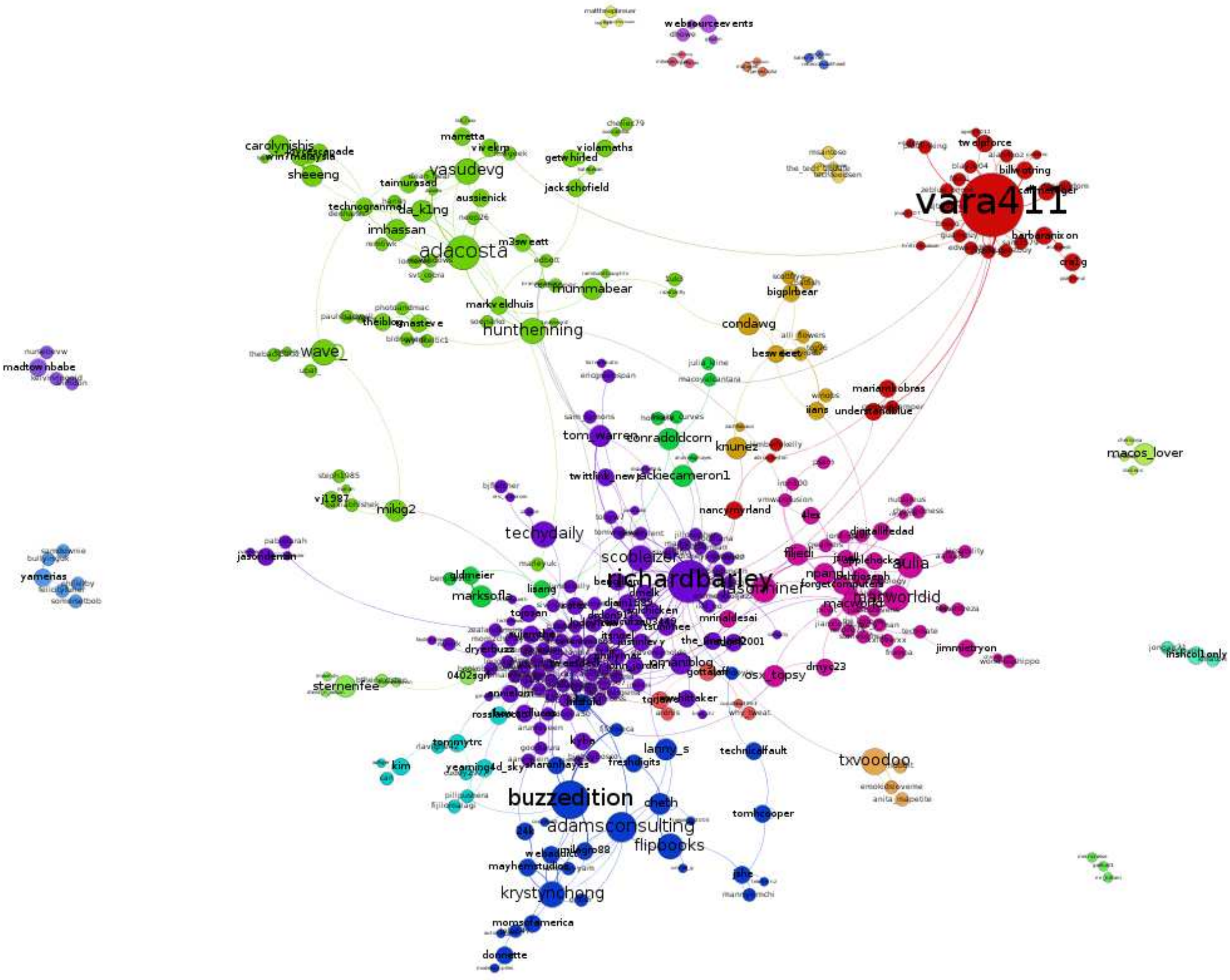


Figure 3.7. Technology users out-degree mention graph

3.5.2 In-degree in mention graph

In-degree distribution can be employed to identify opinion leaders, experts in a particular topic, product or technology. This characteristic in a user-based message is utilized to invite users into a conversation and also to collaborate with others. Figure 3.8 shows that several opinion leaders can be identified in clusters where their expert knowledge is solicited the most.

The graphs in Figure 3.8 and Figure 3.9 are identical, but the node size in Figure 3.7 is proportional to node out-degree, and Figure 3.8 to node in-degree. Organizations can employ this characteristic to identify target users for enhanced and efficient information propagation, as well as for effective social media campaigns.

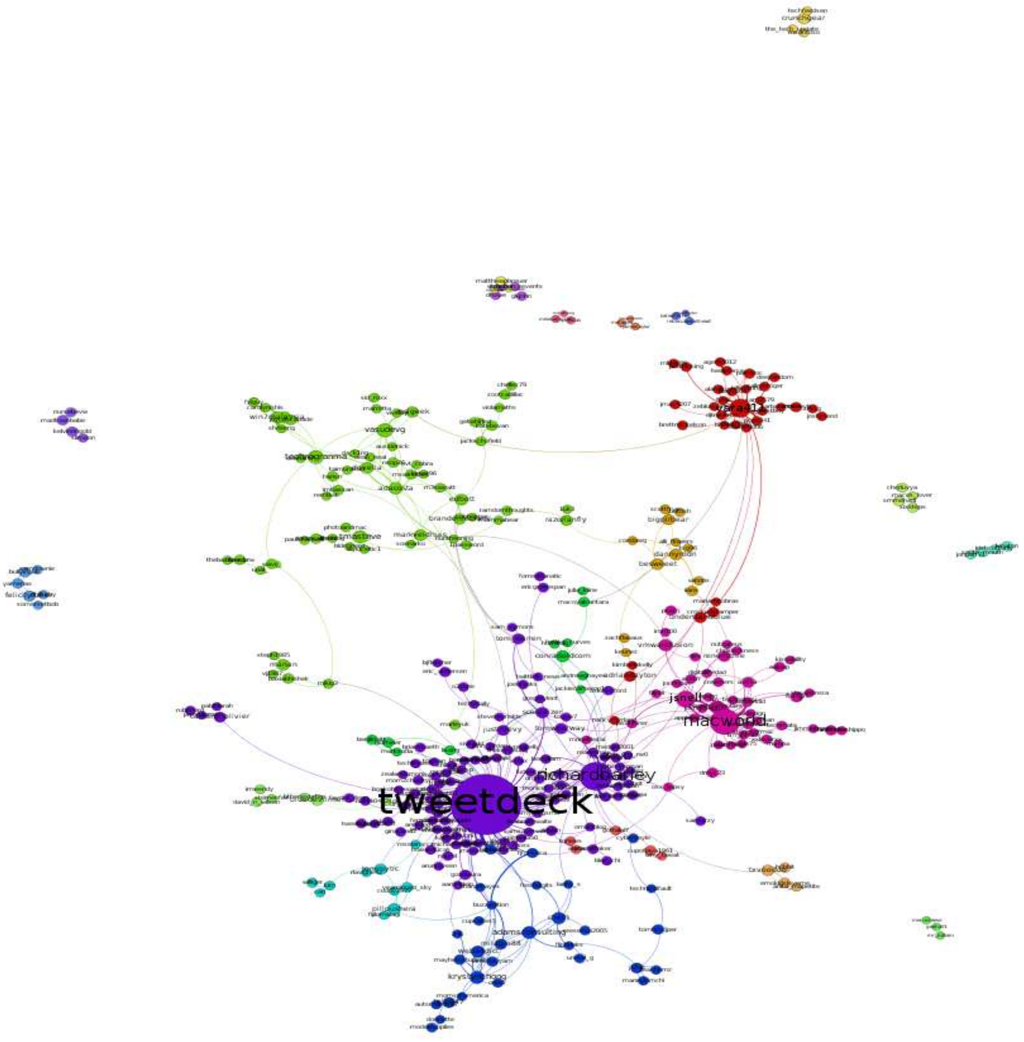


Figure 3.8: Technology users in-degree mention graph

Chapter 4

Proposed approach

This chapter describes the proposed approach to modeling interactions and link prediction of conversations in social networks and Interaction Sentiment Weighting. Thereafter, the experiments, experiment settings, results and discussions are presented.

Interaction Sentiment Weighting can provide insight into conversation polarity of associated users, which can be utilized to model collaborations in social networks. The use of polarities in general, does not apply only to Twitter sentiment propagation dynamics. This approach has potential applications to email messages behavior tracking, consumer behavior and opinion tracking in weblogs. Interaction Sentiment Weighting of users or topics can be employed to better understand the dynamic behavior of social network users or topics. Furthermore, this approach can be crucial in business in order to derive knowledge on how employees in an organisation interact with customers through in-house information technology management systems.

4.1 Interaction Sentiment Weighting (ISW)

In this thesis, we propose an ISW that focuses only on conversations of users in the online social network (OSN) Twitter. Twitter conversations are facilitated by `@` sign, thus conversations of a user facilitated by `@`-mention can be represented by :

$$M_u = \{m_1, m_2, m_3, ...\},$$

where M_u is a set of `@`-mentions of a user u directed at other users in a conversation. M_u consists of sentiment expressed by a user in a conversation, which can be classified as positive, negative or neutral.

Interaction Polarities Modeling that comprises of positive and negative sentiments can be an effective tool for modeling the dynamics of a user, a topic or an entity (e.g., a company) in a social network. Therefore, Interaction Polarity Modeling (IPM) for sets of conversations can be represented by $S(M)$ as denoted below, where M is a set of messages

:

$$S\left(M\right)=\left(\frac{Pos\left(M\right)-Neg\left(M\right)}{Pos\left(M\right)+Neg\left(M\right)}\right) \tag{4.1}$$

$Pos(M)$ and Neg(M) are the numbers of positive and negative mentions [28],respectively, written by a user partaking in a set of conversations and $S\left(M\right)\in[-1,1]$. A high value of $S\left(M\right)$ indicates that a user is inclined to interact in conversations that are positive; conversely, a low value indicates that a user is more often engaged in a conversation that is negative.

Objective or neutral tweets are interactions that do not contain a sentiment or an opinion. These can be fundamental in identifying news Twitter handles, as objective or neutral tweets are tweets that contain headlines that are normally seen in the headers of a newspaper. $Neutral(M)$ is the sum of objective or neutral conversations written by a user partaking in a set of interactions. Objective or neutral tweets of a set of interactions of a user can be modeled by $Obj(M)$ denoted below:

$$Obj\left(M\right)=\left(\frac{Neutral\left(M\right)}{Pos\left(M\right)+Neg\left(M\right)+Neutral\left(M\right)}\right) \tag{4.2}$$

Using these definitions, the ISW function is defined below. The ISW assumes that a user interactions falls within one of the three categories; positive negative or neutral:

$$ISW=\left\{S\left(M\right),Obj\left(M\right)\right\}. \tag{4.3}$$

The ISW comprises Interaction Polarity Modeling and the objective interactions in an OSN such as Twitter and, at the same time, provides a measure to characterize conversation dynamics and the evolution of a network.

4.2 Social profiling

The homophily effect states that similarity breeds connection and individuals are more likely to interact with similar people that share common opinions or interest [44]. This concept, when explored on an OSN, can reveal fascinating characteristics of social users. The analysis of user interaction provides rich data to model social user interest, opinions, and interpersonal characteristics to enhance information propagation and personalized recommendations (a user's interest is represented by the contents of his or her tweets in conversations). A user is similar to another user if they share the same opinion on the same topic (and sentiment) in most interactions. This concept is consistent with the principle of homophily that postulates that a contact or interaction between similar individuals occurs at a greater rate than among dissimilar individuals. A social profile of a user based on his interactions is represented by :

$$I_U=\left\{S\left(M_u\right),Obj\left(M_u\right)\right\}, \tag{4.4}$$

where $S(M_u)$ represents a user sentiment polarities in conversations and $Obj(M_u)$ is the set of objective sentiments in a user's conversations. $S(M_u)$ can be derived by using Definition 4.1 and monitoring a user's interactions and their underlying sentiments.

4.3 Implementation

The implementation of the data preparation, experiments were performed in Python 2.7.10 and 3.4.3. Some of the algorithms are presented in Appendix A. The programming environment used is Linux running the Ubuntu operating system. Pandas version 0.16.2, a Python library, is used for data analysis and handling of data structures. NetworkX and Gephi are used to analyze the network. The NLTK Python library was used for natural language preprocessing steps (i.e., the removal of stop words). Scikit-learn Python package, a collection of machine learning algorithms for the Python programming language, was also employed in the experiments.

4.4 Experiments

In this section, experiments are conducted to validate the ISW. The section starts with the description the sentiment classifier, thereafter, the experimental settings, @-mentions conversation threshold used to develop the ISW features employed in the experiments and processing time are described.

Each of the experiments comprised of a training stage and a test stage. In the training stage the ISW contained the users labelled edges (links). The testing stage, the edges (links) are removed while the future edges are predicted based on the ISW features.

For Interaction Sentiment Modeling, a sentiment classification system was developed to derive sentiments for the ISW. The sentiment classification system was developed utilising training and validation data-sets created by Alec Go, Richa Bhayani and Lei Huang [23]. The machine learning algorithms utilized for the experiments to develop an effective sentiment classification system and a model are Multinomial Naive Bayes, Support Vector Machine (SVM) and Logistic Regression. The performance of these algorithms is presented below.

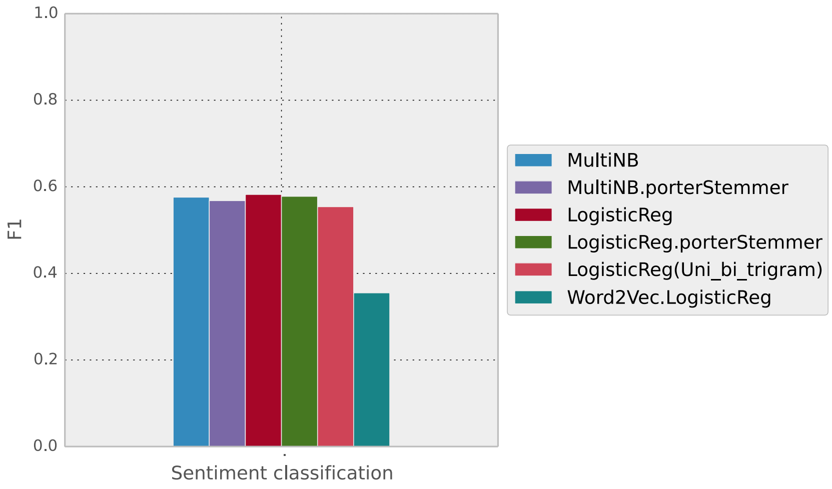


Figure 4.1: F-score sentiment classifiers

The description of the classifiers F-measures are presented in Figure 4.1. The figure shows that MultiNB (Multinomial Naive Bayes) without Porter's stemming (described in Chapter 3) outperformed Multinomial Naive Bayes with Porter stemming by 1.41%. MultiNB and MultiNB.porterStemming shows an almost similar performance without any significant improvement between the two. In addition, Logistic Regression (LogisticReg) without Porters stemming outperformed Logistic Regression with Porter's stemming by 0.7%. The same behavior can be seen for LogisticReg and LogisticReg.porterStemming has an almost similar performance. Investigation with bigrams and trigrams (bags of word sequences approach) shows that Logistic Regression with bigrams and trigrams did not yield a significant improvement to the classifier. Furthermore, Word2Vec algorithm with Logistic Regression had the lowest performance with an F-measure of 0.355 compared to other approaches. Logistic Regression (LogisticReg) had the best overall performance with the highest F-measure above the baseline reported in [1] by 3% and was, therefore, used for further analysis.

4.4.1 Interaction Polarity Modeling (IPM)

In this section Interaction Polarity Modeling (IPM) is presented and the sentiment dynamics of Twitter real-world data is discussed.

Interaction Polarity Modeling (Equation 4.1), one of the features of ISW, is used to model conversations consisting of positive and negative sentiments. IPM provides an indicator to interaction polarities of social users, in order to identify possible negative and positive interaction. Furthermore, IPM is utilized to derive sentiment dynamics of interactions on Twitter. The figures in this section, $S(M)$ indicates Interaction Polarity Modeling (IPM).

Figures 4.2, 4.3 and 4.4, IPM (Equation 4.1) is utilized to extract sentiment dynamics of topics and subtopics found in the Twitter real-world data used in the research (Table 3.3). The topics are in three categories: Sports, Technology and Politics.

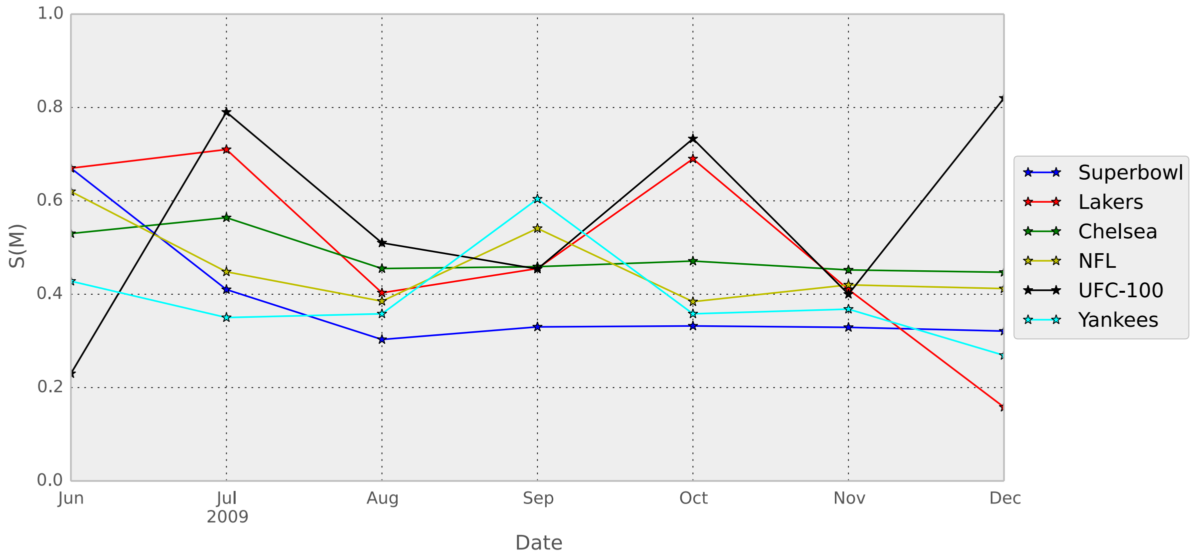


Figure 4.2: Topic-based sentiment dynamics distribution of sport

Figure 4.2, shows the different Interaction Polarity Modeling plots for Sport partitioned topics (Table 3.3). The sports IPM plot shows that Lakers (a basketball team) had a spike of IPM values in the month of June when they won the 63rd edition of the NBA championship series that was played between 4 June and 14 June, which probably indicates fans were positive (pleased) about the team's performance[1]. Another spike in IPM values for the Lakers was in the month of October when the Lakers played their first pre-season game of the season against the Golden State Warriors on 7 October. Furthermore, the IPM value for the Chelsea Football Club shows fans of club were probably pleased about the team's performance from June to December as their club had won the Premier League and FA (Football Association) Cup[2]. The IPM values for UFC-100, a mixed martial arts event, in Figure 4.2, held by the Ultimate Fighting Championship (UFC), the IPM values in the month of 11 July show that the fight attracted a lot of interest, which was named the event of the year[3].

[1] https://en.wikipedia.org/wiki/2009_NBA_Finals

[2] https://en.wikipedia.org/wiki/2009-10_Chelsea_F.C._season

[3] https://en.wikipedia.org/wiki/UFC_100

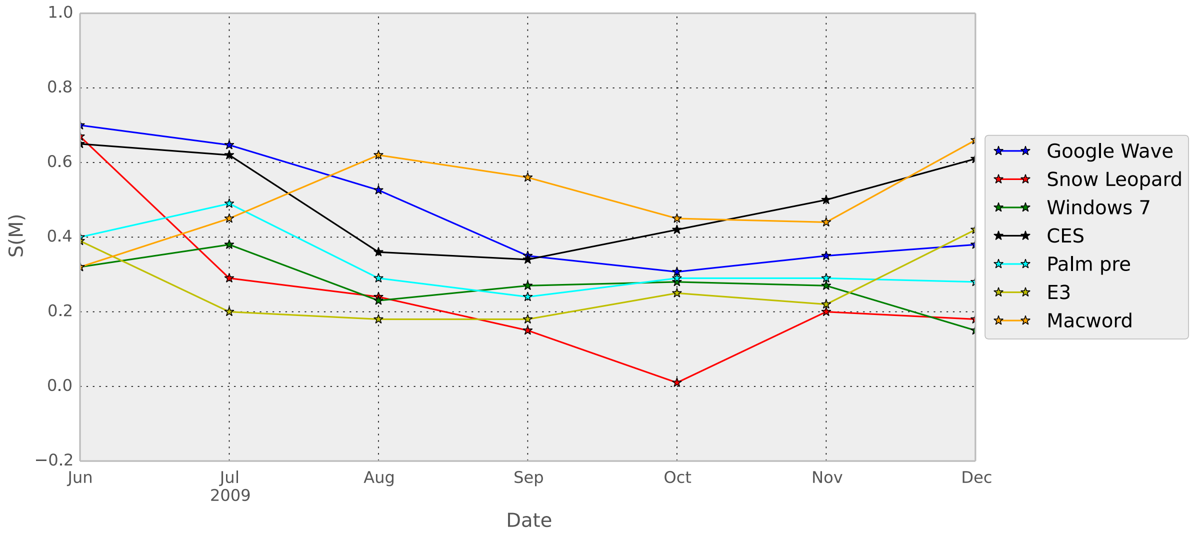


Figure 4.3: Topic-based sentiment dynamics distribution of technology

Figure 4.3 shows the different IPM plots for Technology partitioned topics (Table 3.3). The IPM values for Google Wave, a real time messaging platform that was unveiled in May 2009, shows that in June users were okay with the product; however, there was a subsequent decline in the IPM values which probably indicates that users became unhappy with the product. Google discontinued the development of the product[4]. The IPM values for Macworld, a Macintosh product review and buying advice from Apple experts [5] shows that social users tend to be pleased with the news and tips from Apple experts in the month of August. The Figure 4.3 also shows IPM values of Snow Leopard, the seventh major release of Mac OS X, Apple's desktop and server operating system for Macintosh computer, which was publicly unveiled on 8 June, 2009[6]. The figure indicates a gradual decrease of the IPM values on Snow Leopard from June to December, which probably can be attributed to a lot of enthusiasm for the unveiling in the first week of June; thereafter, the enthusiasm gradually waned.

[4] http://techcrunch.com/2010/08/04/wave-goodbye-to-google-wave/

[5] http://www.macworld.com/

[6] https://en.wikipedia.org/wiki/Mac_OS_X_Snow_Leopard

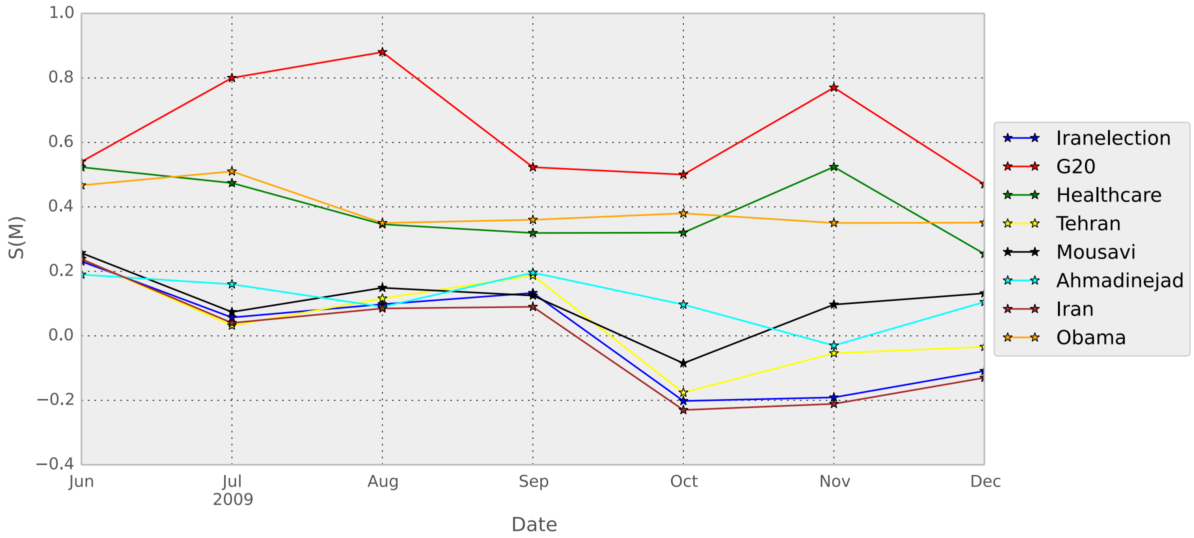


Figure 4.4: Topic-based sentiment dynamics distribution of politics

In Figure 4.4 shows the different IPM plots for politics partitioned topics (Table 3.3). IPM values for the Iranelection relating to elections held on 12 June 2009 with Ahmadinejad running against Mousavi and others, shows negative IPM values which probably correlates with the protest by millions around the world about their disagreement with the election results[7]. Here, social network played an instrumental role, especially Twitter, which was the central-online gathering site during the protests and the organization many of the 2009 Iranian election protests[8]. The IPM values for Iran capital Tehran, where intense 2009 Iranian protests were held [9], the plot shows negative IPM values which probably correlates to the violence in the city. The IPM values were obtained for the 2009 G-20 Pittsburgh summit, September 2009, which involved the third meeting for heads of states to discuss the financial markets and the world economy[10]. Protest were reported to have occurred during the G20 summit and people were wrongfully arrested. The negative values of IPM during the month of September can probably be attributed to protests around the G-20 summit in Pittsburgh. The IPM values for Obama and healthcare are positive, which indicates that social users were probably happy (pleased) with Obama and healthcare from June to December 2009.

[7] https://en.wikipedia.org/wiki/Iranian_presidential_election,_2009

[8] https://en.wikipedia.org/wiki/Internet_activism_during_the_2009_Iranian_election_protests

[9] http://www.nytimes.com/2009/06/14/world/middleeast/14iran.html

[10] https://en.wikipedia.org/wiki/2009_G-20_Pittsburgh_summit

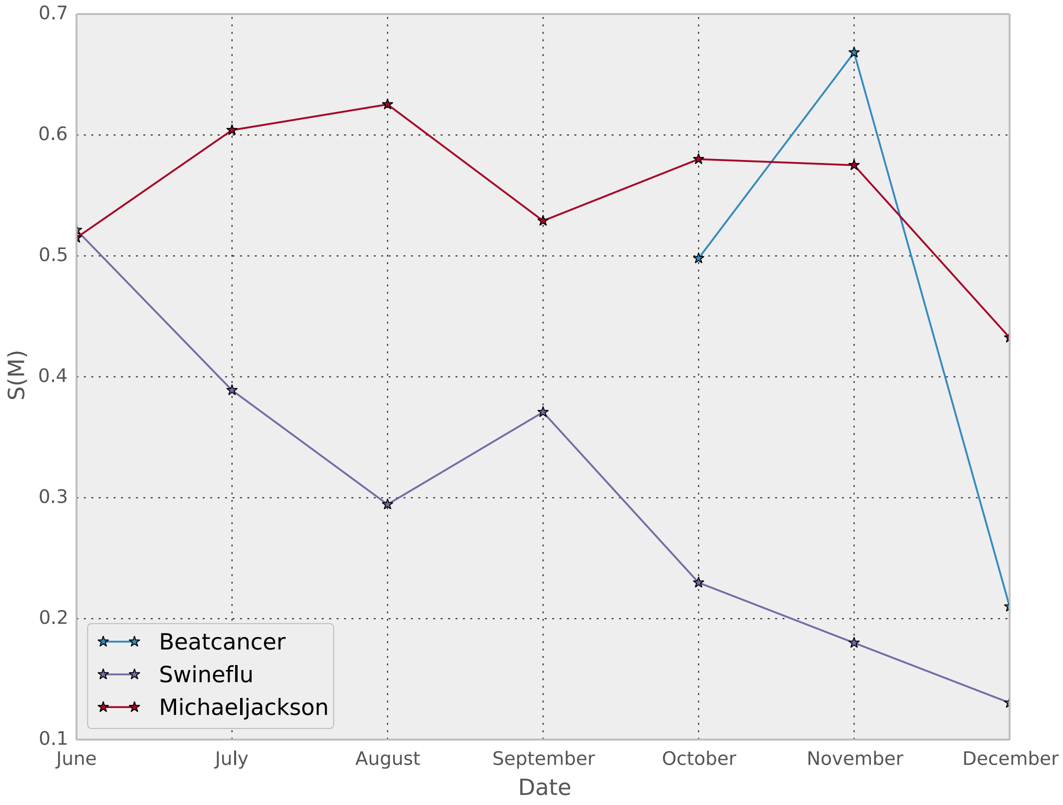


Figure 4.5: Sentiment dynamics distribution of notable-events

IPM dynamics for three major events in 2009 based on the whole Twitter mention graph are presented in Figure 4.5. One event, *beatcancer*, was a social media campaign for cancer charities, whose goal was to set a Guinness world record for the most social mentions in 24 hours on the 16^{th} October, 2009[11]. The promise was that for every tweet containing the #beatcancer hashtag, Ebay/Paypal and Millercolors would be donating 1 cent to cancer research, thereby raising significant money for several cancer charities. Figure 4.6 shows that #beatcancer had high momentum in the month of October having an an IPM value of 0.668, which possibly indicates that most of the social users were positive about beating cancer, thereafter the IPM value declines. The campaign succeeded in achieving a Guinness world record.

Figure 4.5 shows swineflu dynamics, during an influenza pandemic that was announced in late April 2009 by the World Health Organization (WHO), which was declared to be the first ever "public health emergency of international concern"[12]. The swineflu pandemic

[11] http://mashable.com/2009/10/19/beatcancer-sets-record/
[12] https://en.wikipedia.org/wiki/2009_flu_pandemic

began to diminish in November 2009 and the number of cases was in steep decline when the WHO announced an end to the pandemic.

The IPM for Michael Jackson are presented in Figure 4.5. The 'king of pop', Michael Jackson, passed away on 25 June, 2009[13]. In the figure, the IPM dynamics are investigated. Jackson's death triggered grief around the world, creating unprecedented surges of internet traffic and causing sales of his music and that of the Jackson 5 to increase dramatically[14]. The figure possibly indicates that despite Michael Jackson's death, the world (social users) had positive messages to tweet about his musical achievements.

4.4.2 Experiment settings

Social networks are dynamic and grow via the addition of new nodes and edges; therefore, it is not practical to seek predictions for validation (test) interval edges whose endpoints are not in the training interval [37]. In this experiment the training interval of the mention graph is defined as the period t and the test interval is the node and edges of time $t+1$, which are to be predicted. A threshold was used to produce a validation dataset. A minimum of six interactions in the timeline of social users for the partitioned topics (Technology, Sports and Politics in Sections 3.4 and 3.5) was used to create validation data for the time interval $t+1$ to verify the approach. After data preprocessing and applying link prediction theory described in Section 2.7, the sampled data comprised of 21,818 interactions. The data preprocessing schematics is presented in Appendix B. Experiments were performed by utilizing 21,818 interactions data. Furthermore, to prevent overfitting, cross-validation used (as explained in Chapter 2).

Processing time

The processing time (the time it takes to perform or run an experiment) was evaluated using the Python library Timeit[15]module. The training and testing stages processing times were measured separately and are presented in Appendix B.

4.5 Results and discussions

In this section, we will start by presenting experiments conducted using IPM with regards to link prediction. Thereafter, we will present the experiments conducted on ISW.

The evaluation and performance measures used to evaluate the results are: precision, recall and the F-measure (F1 score). In Figure 4.6, we investigate the IPM as the number of training data-sets increases. In the Figure 4.6 below SM indicates IPM.

[13] https://en.wikipedia.org/wiki/Michael_Jackson

[14] https://en.wikipedia.org/wiki/Death_of_Michael_Jackson

[15] https://docs.python.org/2/library/timeit.html

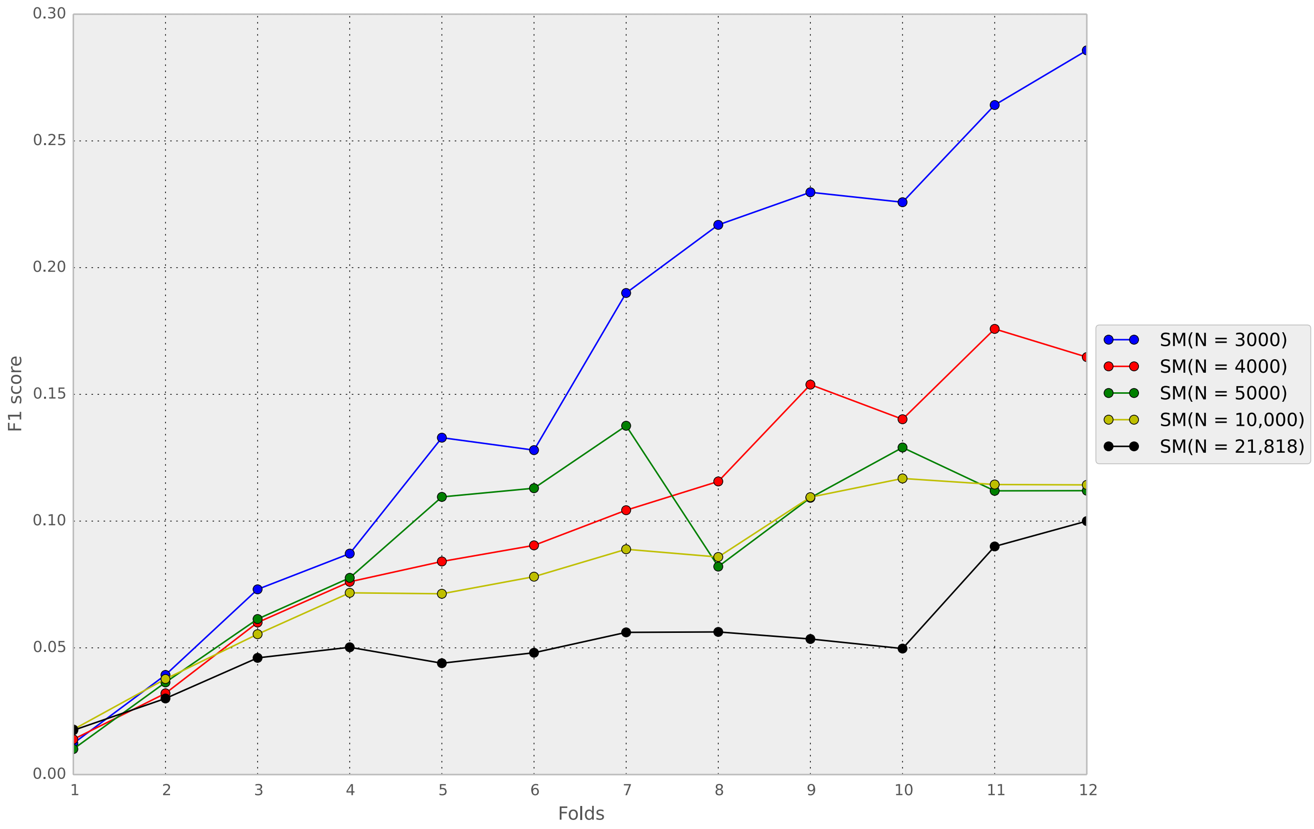


Figure 4.6: Cross-validation F-measure distribution

Figure 4.6 shows the results of the classifier trained using different values of K, it indicates that classifiers trained with more folds perform better. The figure further shows the best F-measure is when the number of nodes is 2,000 with an F-measure of 0.44 in the 12^{th} fold. Furthermore, as the number of nodes increases the performance of model improves with each fold. This behavior indicates the IPM F-measure increases as the number of training sets and the number of folds increases.

The F-measure of the IPM (Equation 4.1) predictor based on different number of nodes and edges is shown in Figure 4.6, indicates that cross-validation enhances the performance of the IPM predictor as the performance gets better after each fold in almost all cases.

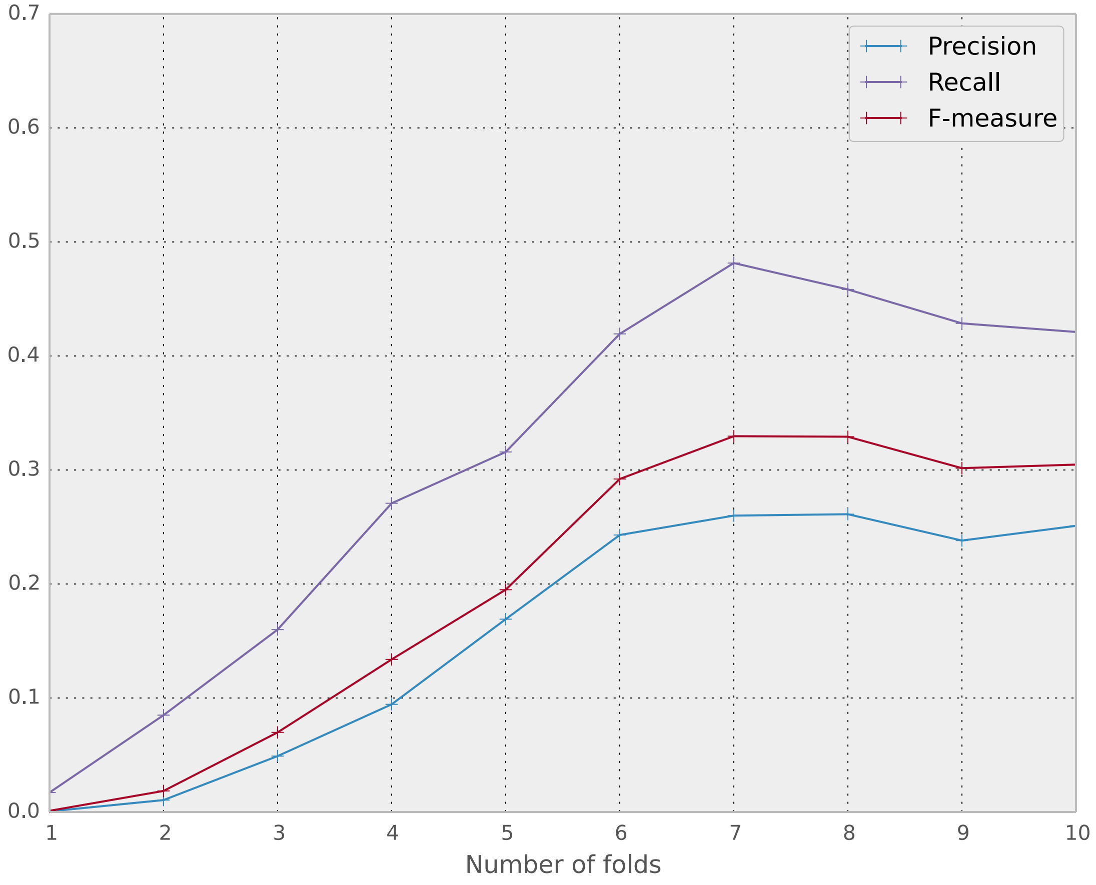


Figure 4.7: Precision/recall/F-measure distribution

Figure 4.7 presents the precision, recall and F-measure distributions of IPM. The figure shows the results of the classifier trained using different numbers of K and the precision, recall and F-measure distribution. The precision, recall and F-measure tend to be optimal when K is 6 or 7 and when the interaction data is 1,138 chosen manually.

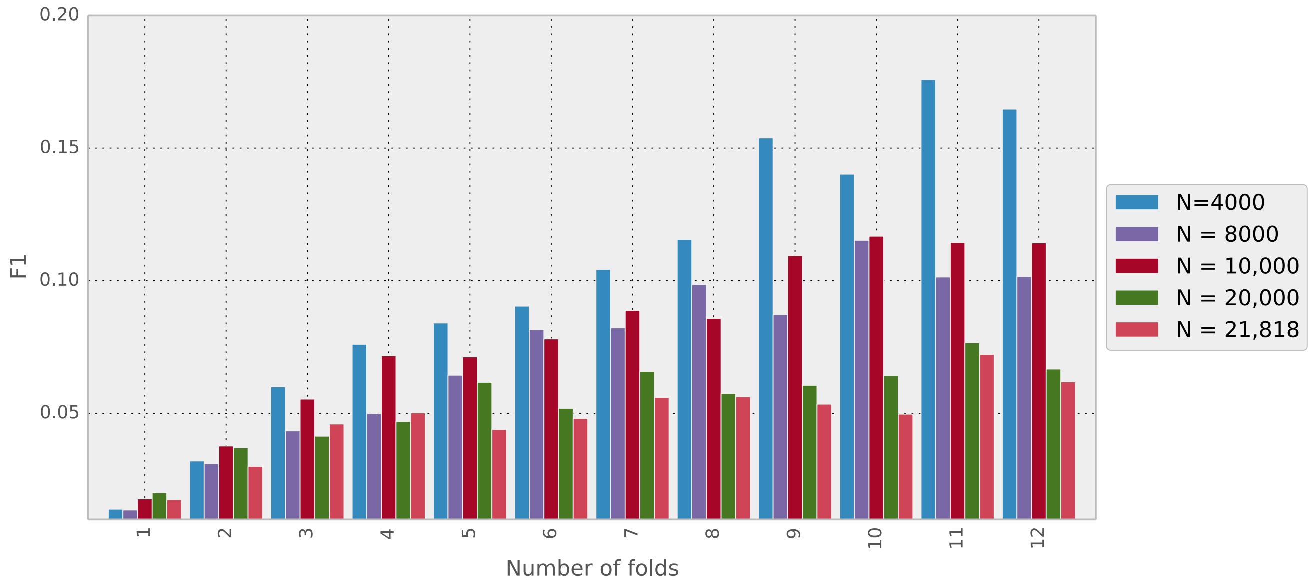


Figure 4.8: F-measure distribution of training data

In Figure 4.8, the IPM, the first round or first fold of cross-validation, the model performs its worst with a value close to zero. However, multiple rounds of cross-validation using different partitions the resulting F-measure increases to approximately 0.17 for 4,000 nodes, using more fold. Utilizing cross-validation, each link is used for prediction exactly once, which reduces statistical bias (an error you cannot correct by repeating the experiment many times and averaging together the results).

Interaction Polarity Modeling and other link prediction approaches

Comparison of the IPM predictor and random predictors is investigated in this subsection. Random prediction is used as the baseline of the research. Random prediction is prediction based on known data; that is, known data is used to predict a future event through a random event generator. According to May et al. [43], random prediction was used for prediction with the intention of measuring global consciousness (a research exploring possible interactions of human consciousness and emotions with the physical system).

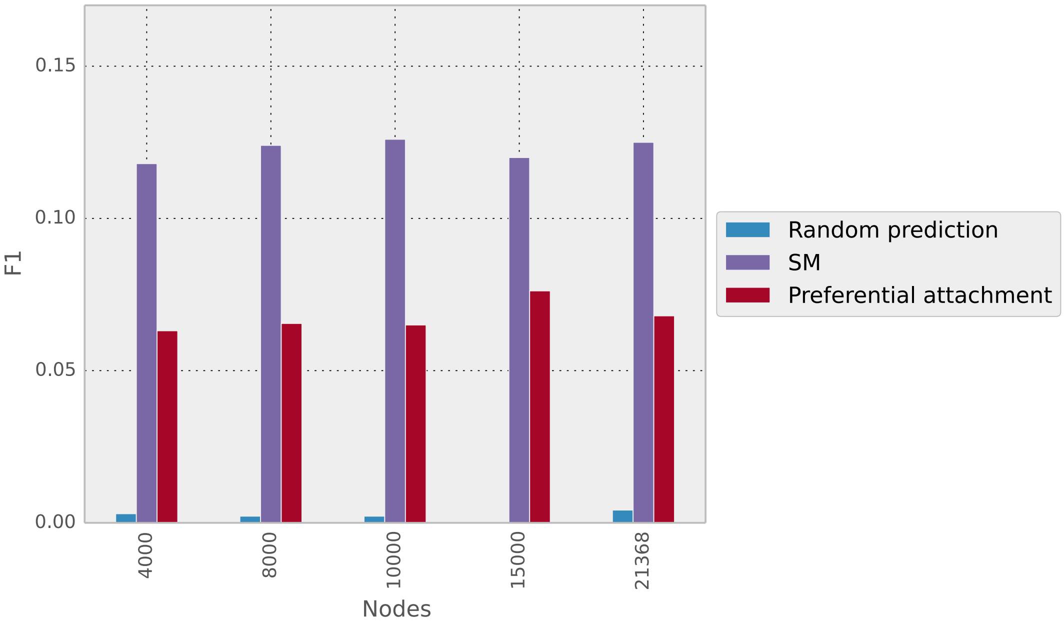


Figure 4.9: Comparison of Interaction Polarity Modeling and Random prediction

In Figure 4.9, the number of future links to predict was set to 450 edges and the training data-set sample were increased while IPM and random predictor experiments were conducted. IPM performed better than random predictor in all of the experiments conducted. The Figure 4.9 indicates that IPM (SM) predictor outperformed the random predictor by a factor of 50 to 60.

The preferential attachment approach was implemented and its performance compared with Interaction Polarity Modeling approach (Figure 4.9). Preferential attachment has attracted a lot of interest as a model of the growth of networks [37]. The underlying assumption of the preferential attachment model is that the probability that a new edge involves node x is proportional to $|\Gamma x|$ (the current number of neighbors of x). Figure 4.9 indicates that the IPM approach outperforms the preferential attachment approach by 80%.

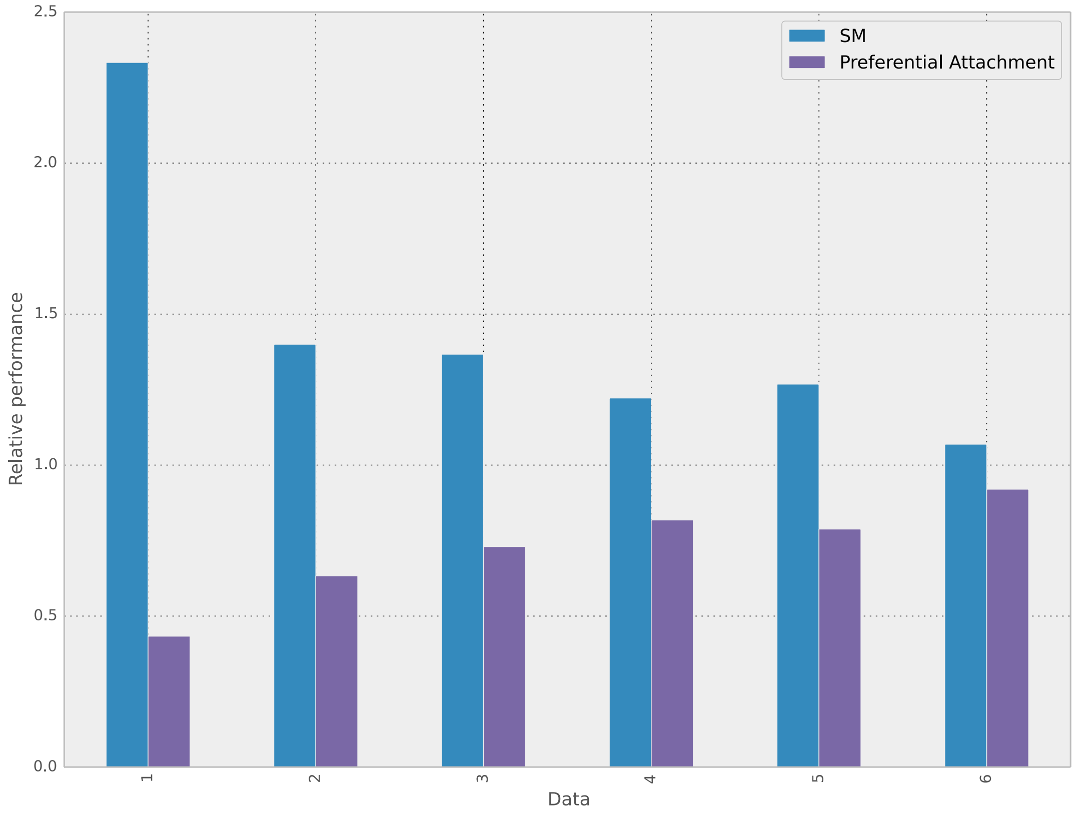


Figure 4.10: Relative performance of SM and Preferential attachment

Figure 4.10 shows the relative performance of SM (IPM) and preferential attachment. This plot provides a view of the performance of the IPM relative to the preferential attachment approach. The x-axis values consisting of 1,2,3,4,5 and 6 represents data distribution of the consisting of nodes [4,000],[6,000], [8,000],[10,000], [15,000] and [21,818], respectively. The figure show that the IPM approach performed better than preferential attachment in all cases in modeling network evolution with a relative average performance of 1.44.

ISW and other link prediction approaches

In this subsection, we investigate ISW performance with respect to link prediction. We start with the investigation of class imbalance of objective tweets with different representation models for text classification. Thereafter, the relative average performance of ISW versus preferential attachment and a random predictor are presented.

The sentiment classification system was enriched with 5,389 objective (neutral) tweets of multiple brands. The class imbalance was explored as well as the effect on the model performance.

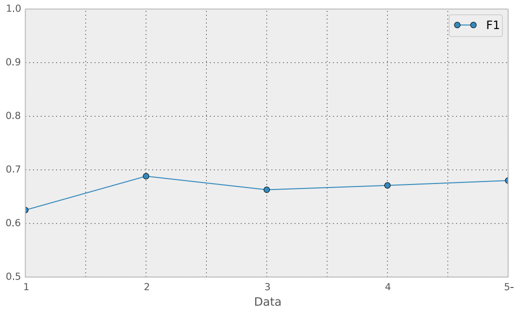


Figure 4.11: Bags of word approach F-measure label class distibution

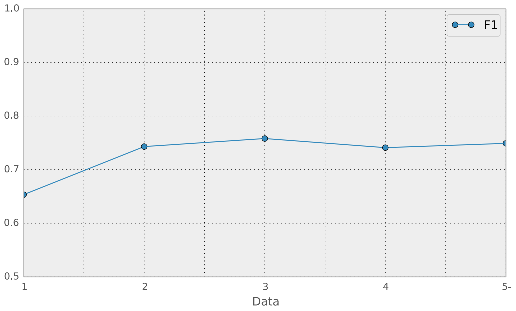


Figure 4.12: Bags of word sequences approach F-measure class label distibution

Figure 4.11 and Figure 4.12 show the F-measures of two representation models for text classification, the bags of word approach and the bags of word sequences approach. This plot provides a view of the performance of the classifier with imbalance classes. The x-axis values consisting of 1,2,3,4 and 5 represent the label class distribution of the training data-set neutral, negative and positive consisting of [5,000-5,000-5,000],[5,000-10,000-10,000], [5,000-20,000-20,0000],[5,000-30,000-30,000],[5000-40,000-40,000] respectively. The figures show that with an increase in the size of the data-set used to learn the models, the model performance also increases. A balanced class model of the bags of word sequences approach was selected for the remaining experiments to avoid class bias towards major classes and very poor classification rates on minor classes.

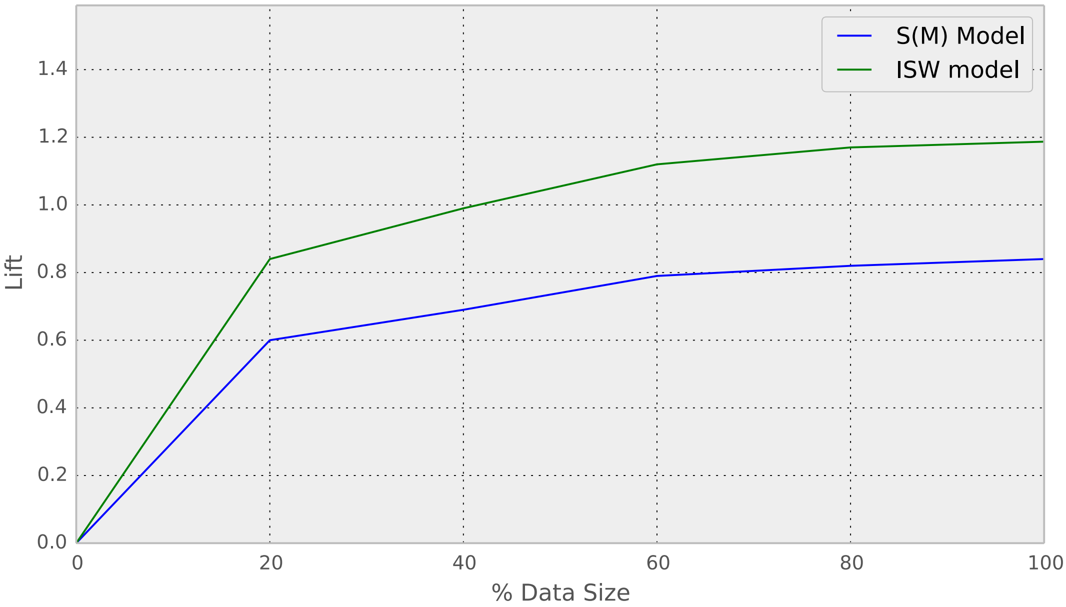


Figure 4.13: Lift chart

.

Figure 4.13 presents the lift chart of the ISW model compared with the Interaction Polarity Model. The data sampled with interaction threshold of more than one conversation for $t+1$ for Technology data (Table 3.5). After preprocessing and applying the link prediction theory described in Section 2.7, the data comprised of 11,047 interactions. The model was validated with 306 interactions for time $t+1$ and 10,741 interactions for training the model . We can see from the figure that the ISW model outperformed the Interaction Polarity Model.

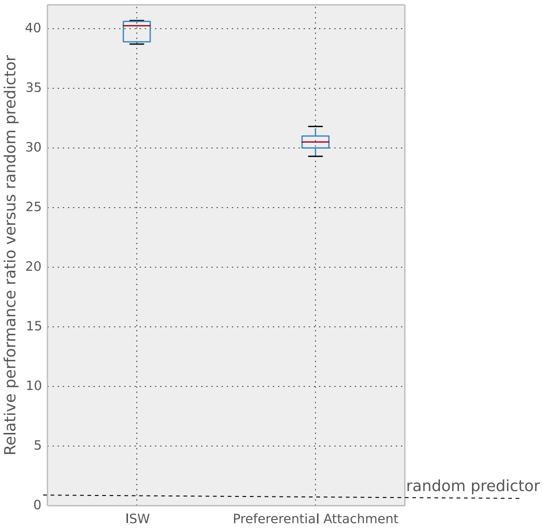


Figure 4.14: Relative average performance of ISW and preferential attachment versus random predictor

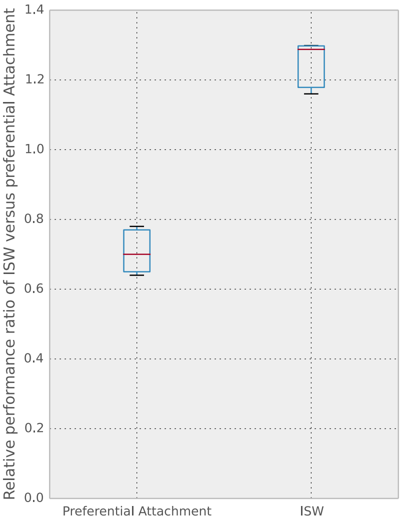


Figure 4.15: Relative average performance of ISW versus preferential attachment

Figure 4.14 shows the relative average performance ratios of ISW and preferential attachment versus the random predictor. The displayed values show that ISW outperformed random prediction by a ratio of 40 to 1. Figure 4.15 shows the relative average performance ratio of ISW versus the preferential attachment. The displayed values show that ISW outperformed preferential attachment by a ratio of 1.29 to 0.78.

Chapter 5

Conclusions and future work

This chapter presents the conclusions for the research study and then gives suggestions on directions for future work.

5.1 Conclusion

The research was started with the goal of optimization of conversation in social networks (e.g., Twitter) given a collection of user sentiments and the underlying topics. The concept of homophily (when two people share more interests or opinion, they have a higher probability to be connected and interact with each other in a social network) is explored, to optimize conversation recommendation. This concept is used to develop the Interaction Sentiment Weighting (ISW) function, using social user's interactions, the mention graph, the sentiment and the underlying topic in a social network. The result shows that ISW improves conversation recommendation with greater performance than the baseline.

We can deduce from the research results that the ISW approach outperforms the random predictor and preferential attachment (using Twitter dataset), suggesting that there is indeed useful information contained in the interactions of social users in a network. Furthermore, the ISW approach is shown to be effective in inferring the sentiment dynamics of Twitter real-world data, which can be effective in modeling the wave of interactions to characterize a user, a product or an entity (e.g., a company). In addition, the data-set imperfections shows that working with real-world data-sets can be challenging, as we have presented the data preprocessing steps and discovered that data-sets are sparse as shown in the data characteristics.

5.2 Future work

Beyond this thesis, there are a lot of interesting areas that require improvement. The Interaction Sentiment Weighting (ISW) has been shown to be an effective approach to model network evolution with a performance greater than the random approach and the preferential attachment approach for Twitter real-life data. However, the current study gives only the results for one data-set. The behavior for other data-sets with different characteristics might be very different.

Another direction for future work, from a more general data mining perspective on social networks, is the use of more information about the social user. In this study, the only information that was known about the social user is the user's @-mentions, the messages, and the time the messages were sent. However, much more information can be utilized, such as the user's followers and followees.

Furthermore, another direction for future work is link prediction in multi-dimensional networks, where links could have different meanings. For example, a social network may consist of positive and negative links, respectively pointing to friends and foes or trusted and distrusted peers. The prediction of the existence and the sign of links as either positive or negative has not been well studied [41]. The ISW approach can possibly be employed and may be an effective approach to predict signs of links that could provide useful insights.

Appendix A

Algorithm pseudocode

The pseudocode is written to some extent in a Pythonic way. A *dict* also known as dictionary, which is a Python data type and a container consisting of a set of $key : value$ pairs, where each of the keys are unique within one dictionary. A pair of empty braces { } creates an empty dictionary and $dict[key]$ will return all values given the key. The major operations of a *dict* is storing a value with a key and extracting the value with the key. A $dict()$ constructor builds dictionaries directly from sequences of key-value pairs.

Algorithm 1 S(m) algorithm

```
procedure S(M)–PROCEDURE
    filename = filename
    positive = 0
    negative = 4
    filedictionary = dict()
    order = []
    count = { }
    with open(filename,'r') as file do
    for each in line file do
        split = line.split(",")
        if (split[0] not in fileDictionary) then
            fileDictionary[split[0]] = []
            counts[split[0]] = [0,0]
        end if
        order.append(split[0])
        fileDictionary[split[0]].append(split[1])
    end for
    with open(filename,'w') as file do
    file.write to file
    for  key in order do
        elem = fileDictionary[key].pop(0)
        if (elem == negative) then
            counts[key][1] += 1
        else if (elem == positve) then
            counts[key][0] += 1
        end if
        negatives =counts[key][0]
        positives = counts[key][1]
        file.write to file
        file.write format(key,elem,(positives-negatives)/(float(positives+negatives)))
    end for
end procedure
```

Algorithm 2 Obj(m) algorithm

```
procedure Obj(m)–procedure
    filename = filename
    positve = 0
    negative = 4
    neutral = 3
    filedictionary = dict()
    order = []
    count = { }
    with open(filename,'r') as file do
    for each in line file do
        split = line.split(",")
        if (split[0] not in fileDictionary) then
            fileDictionary[split[0]] = []
            counts[split[0]] = [0,0]
        end if
        order.append(split[0])
        fileDictionary[split[0]].append(split[1])
    end for
    with open(filename,'w') as file do
    file.write to file
    for for key in order do
        elem = fileDictionary[key].pop(0)
        if (elem == negative) then
            counts[key][1] += 1
        else if (elem == positive) then
            counts[key][0] += 1
        else if (elem == neutral) then
            counts[key][2] += 1
        else
            continue
        end if
        negatives =counts[key][0]
        positives = counts[key][1]
        neutrals = counts[key][2]
        ObjMu = (neutral)/float((positive+negative+neutral))
        file.write to file
        file.write format(key,elem,ObjMu))
    end for
    print("Done")
end procedure
```

Appendix B

Processing time and data preprocessing schematics

The Table B.1 shows the processing time of IPM, ISW, preferential attachment and random prediction implementations for both the training and testing stage.

The table shows a linear processing time, the processing time is directly proportional to the input size, i.e., time grows linearly as input size increases.

Technique	Training stage	Testing stage
ISW (N = 500)	1.29 s	240 ms
ISW (N = 1000)	5.72 s	242 ms
ISW (N = 2000)	32.5 s	242 ms
ISW (N = 3000)	1min 29 s	242 ms
ISW (N = 4000)	3 min 9 s	244 ms
ISW (N = 4653)	5 min 58 s	257 ms
ISW(N = 10,000)	8 min 47s	260 ms
Preferential Att. (N = 500)	2.21 s	232 ms
Preferential Att. (N = 1000)	8.39 s	232 ms
Preferential Att. (N = 2000)	40.3 s	233 ms
Preferential Att. (N = 3000)	1min 47 s	231 ms
Preferential Att. (N = 4000)	3 min 21 s	231 ms
Preferential Att. (N = 4653)	4 min 43 s	233 ms
Preferential Att. (N=10,000)	7 min 45 s	235 ms
SM (N=4000)	3 min 23 s	231 ms
SM (N=6000)	5 min 12 s	233 ms
SM (N=8000)	6 min 34 s	236 ms
SM (N=10,000)	7 min 41s	239 ms
SM (N=21,814)	12 min 53 s	244 ms
Random pred.(N=500)	16.8 µs	2.26 ms
Random pred.(N=1000)	17.7 µs	2.27 ms
Random pred.(N=2000)	18.3 µs	2.28 ms
Random pred.(N=3000)	19 µs	2.27 ms
Random pred.(N=4000)	20.9 µs	2.28 ms
Random pred.(N=4653)	22.5 µs	2.28 ms

Table B.1: Processing time for the training and test stages

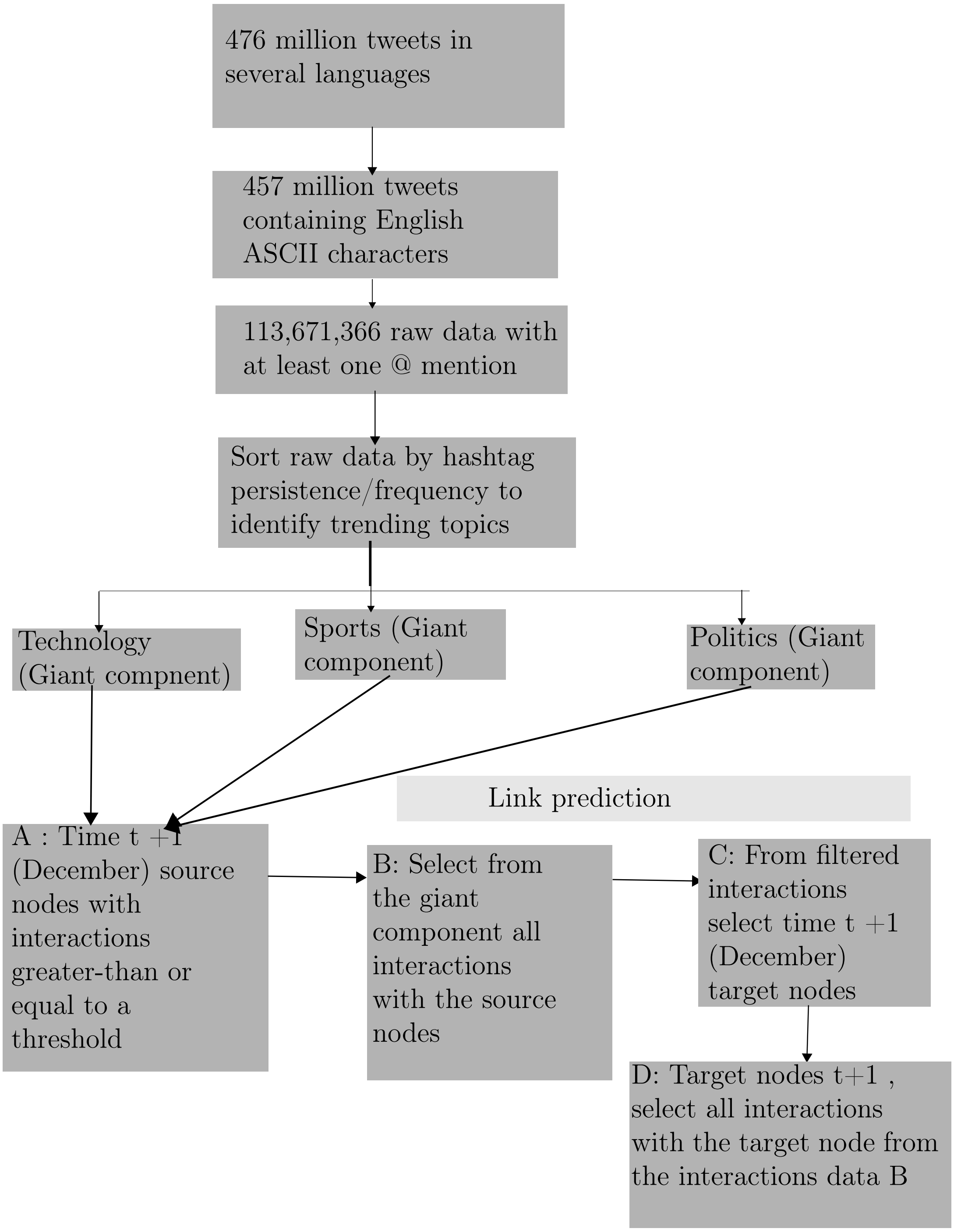

476 million tweets in several languages
457 million tweets containing English ASCII characters
113,671,366 raw data with at least one @ mention
Sort raw data by hashtag persistence/frequency to identify trending topics
Technology (Giant compnent)
Sports (Giant component)
Politics (Giant component)
Link prediction
A : Time t +1 (December) source nodes with interactions greater-than or equal to a threshold
B: Select from the giant component all interactions with the source nodes
C: From filtered interactions select time t +1 (December) target nodes
D: Target nodes t+1 , select all interactions with the target node from the interactions data B

Figure B.1: Data preprocessing schematics

Figure B.1 shows the schematics of data preparation used in the experiments. The figure presents the steps used to preprocess 476 million tweets to the interactions data used in the research experiments and results presented in chapter 4.

List of Notations

$F - measure$ The harmonic mean between precision and recall

I_u The social user profile based on his/her interactions

ISW The Interaction Sentiment Weighting (ISW) based on a user interaction that falls within three categories: positive, negative or neutral

M_u The set of mentions of a user

$Neg(M)$ Count of negative mentions by a user in a set of interaction

$Neutral(M)$ Count of neutral mentions of by a user in a set of interaction

$Obj(M)$ The objectivity sentiment value that charactises a user in a set of interactions

$Pos(M)$ Count of all positive mentions of by a user in a set of messages

$Precision$ The proportion of predicted positives which are actual positive

$Recall$ The proportion of actual positives that are predicted positive

$S(M)$ The Interaction Polarity Modeling value comprising of positive and negative sentiments in a set of messages in an interaction

Bibliography

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (2011), Association for Computational Linguistics, pp. 30–38.

[2] Agrawal, D., Bamieh, B., Budak, C., El Abbadi, A., Flanagin, A., and Patterson, S. Data-driven modeling and analysis of online social networks. In *Web-Age Information Management*. Springer, 2011, pp. 3–17.

[3] Ayodele, T. O. *Machine learning overview*. New Advances in Machine Learning, Yagang Zhang (Ed.), InTech, DOI: 10.5772/9374., 2010.

[4] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM International Conference on Web search and Data mining (WSDM)* (2011), ACM, pp. 65–74.

[5] Bekkerman, R., and Allan, J. Using bigrams in text categorization. *Department of Computer Science, University of Massachusetts, Amherst 1003* (2004), 1–2.

[6] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*, 10 (2008), P10008.

[7] Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1–8.

[8] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)* (Washington DC, USA, May 2010).

[9] Chang, C.-C., and Lin, C.-J. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27.

[10] Chui, M. *The social economy: Unlocking value and productivity through social technologies*. McKinsey, 2013.

[11] Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., and Tucci, G. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research* (2012), ACM, pp. 25–31.

[12] Dang-Xuan, L., and Stieglitz, S. Impact and diffusion of sentiment in political communication-an empirical analysis of political weblogs. In *In Proceedings of the sixth International AAAI Conference on Weblogs and Social Media (ICWSM)* (2012), pp. 427–430.

[13] Deng, L., and Yu, D. Deep learning for signal and information processing. *Microsoft Research Monograph* (2013).

[14] Ding, X., Liu, B., and Yu, P. S. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (2008), ACM, pp. 231–240.

[15] Easley, D., and Kleinberg, J. *Networks, crowds, and markets: Reasoning about a highly connected world.* Cambridge University Press, 2010.

[16] Ellison, N. B., et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication 13*, 1 (2007), 210–230.

[17] Flay, B. R. Evaluation of the development, dissemination and effectiveness of mass media health programming. *Health Education Research 2*, 2 (1987), 123–129.

[18] Fortunato, S. Community detection in graphs. *Physics Reports 486*, 3 (2010), 75–174.

[19] Gabielkov, M., and Legout, A. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM Conference on CoNEXT Student Workshop* (2012), ACM, pp. 19–20.

[20] Getoor, L. Link mining: a new data mining challenge. *ACM SIGKDD Explorations Newsletter 5*, 1 (2003), 84–89.

[21] Getoor, L., and Diehl, C. P. Link mining: a survey. *ACM SIGKDD Explorations Newsletter 7*, 2 (2005), 3–12.

[22] Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., and Tserpes, K. Representation models for text classification: a comparative analysis over three web document types. In *Proceedings of the Second International Conference on Web Intelligence, Mining and Semantics* (2012), ACM, p. 13.

[23] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* (2009), 1–12.

[24] Goutte, C., and Gaussier, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval*. Springer, 2005, pp. 345–359.

[25] Greetham, D. V., and Ward, J. A. Conversations on twitter: Structure, pace, balance.

[26] Guille, A., Hacid, H., Favre, C., and Zighed, D. A. Information diffusion in online social networks: A survey. *ACM SIGMOD Record 42*, 2 (2013), 17–28.

[27] Gupta, N., and Singh, A. A novel strategy for link prediction in social networks. In *Proceedings of the 2014 CoNEXT on Student Workshop* (2014), ACM, pp. 12–14.

[28] Gurini, D. F., Gasparetti, F., Micarelli, A., and Sansonetti, G. A sentiment-based approach to twitter user recommendation. In *RSWeb@ RecSys* (2013).

[29] Gustin, S. Social media sparked, accelerated egypt's revolutionary fire, 2011.

[30] Harbor, R. Infinite Interactions Drives New Values the internet of things meets the internet of people, 2013.

[31] Honey, C., and Herring, S. C. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS'09. Forty-second Hawaii International Conference on* (2009), IEEE, pp. 1–10.

[32] Huang, J. *Performance measures of machine learning.* University of Western Ontario, 2006.

[33] James, J. How much data is created every minute. *Retrieved November 3* (2012).

[34] Jermyn, P., Dixon, M., and Read, B. J. Preparing clean views of data for data mining. *ERCIM Work. on Database Res* (1999), 1–15.

[35] Kim, J., and Yoo, J. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *Social Informatics (SocialInformatics), 2012 International Conference on* (2012), IEEE, pp. 131–136.

[36] Kywe, S. M., Lim, E.-P., and Zhu, F. A survey of recommender systems in twitter. In *Social Informatics*. Springer, 2012, pp. 420–433.

[37] Liben-Nowell, D., and Kleinberg, J. The link-prediction problem for social networks. *Journal of the American society for information science and technology 58*, 7 (2007), 1019–1031.

[38] Lin, C. J., Hsu, C.-W., and Chang, C.-C. A practical guide to support vector classification. *National Taiwan U., www. csie. ntu. edu. tw/ cjlin/papers/ guide/ guide. pdf* (2003).

[39] Liu, L., Tang, J., Han, J., Jiang, M., and Yang, S. Mining topic-level influence in heterogeneous networks. In *Proceedings of the Nineteenth ACM International Conference on Information and Knowledge Management* (2010), ACM, pp. 199–208.

[40] Lu, C., Lam, W., and Zhang, Y. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-sixth AAAI Conference on Artificial Intelligence* (2012).

[41] Lü, L., and Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications 390*, 6 (2011), 1150–1170.

[42] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[43] May, E. C., and Spottiswoode, S. J. P. Global consciousness project: An independent analysis of the 11 september 2001 events, 2001.

[44] McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001), 415–444.

[45] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the Sixteenth International Conference on World Wide Web* (2007), ACM, pp. 171–180.

[46] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (2013), pp. 3111–3119.

[47] Myers, S. A., Sharma, A., Gupta, P., and Lin, J. Information network or social network?: The structure of the twitter follow graph. In *Proceedings of the Companion Publication of the Twenty-third International Conference on World Wide Web Companion* (2014), International World Wide Web Conferences Steering Committee, pp. 493–498.

[48] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. Text classification from labeled and unlabeled documents using em. *Machine learning 39*, 2-3 (2000), 103–134.

[49] Noack, A. *Unified quality measures for clusterings, layouts, and orderings of graphs, and their application as software design criteria.* PhD thesis, Brandenburg University of Technology Cottbus, 2007.

[50] O'Madadhain, J., Hutchins, J., and Smyth, P. Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explorations Newsletter 7*, 2 (2005), 23–30.

[51] Popescul, A., and Ungar, L. H. Statistical relational learning for link prediction. In *IJCAI workshop on Learning Statistical Models from Relational Data* (2003), vol. 2003, Citeseer.

[52] Radovanović, M., and Ivanović, M. Text mining: Approaches and applications. *Novi Sad J. Math 38*, 3 (2008), 227–234.

[53] Rahm, E., and Do, H. H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull. 23*, 4 (2000), 3–13.

[54] Rattigan, M. J., and Jensen, D. The case for anomalous link discovery. *ACM SIGKDD Explorations Newsletter 7*, 2 (2005), 41–47.

[55] Refaeilzadeh, P., Tang, L., and Liu, H. Cross-validation. In *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[56] Romero, D. M., Meeder, B., and Kleinberg, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the Twentieth International Conference on World Wide Web* (2011), ACM, pp. 695–704.

[57] Safro, I., Sanders, P., and Schulz, C. Advanced coarsening schemes for graph partitioning. *Journal of Experimental Algorithmics (JEA) 19* (2015), 2–2.

[58] Saif, H., Fernández, M., He, Y., and Alani, H. On stopwords, filtering and data sparsity for sentiment analysis of twitter.

[59] Sanders, P., and Schulz, C. High quality graph partitioning. *Graph Partitioning and Graph Clustering 588* (2012), 1.

[60] Schein, R., Wilson, K., and Keelan, J. E. *Literature review on effectiveness of the use of social media: a report for peel public health.* [Region of Peel], Peel Public Health, 2010.

[61] Strickland, J. *Predictive Modeling and Analytics.* Lulu. com, 2014.

[62] Su, W.-C. Integrating and mining virtual communities across multiple online social networks: Concepts, approaches and challenges. In *Digital Information and Communication Technology and Applications (DICTAP), 2014 Fourth International Conference* (2014), IEEE, pp. 199–204.

[63] Tan, C., Lee, L., and Pang, B. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438* (2014).

[64] Trung, D. N., and Jung, J. J. Sentiment analysis based on fuzzy propagation in online social networks: a case study on tweetscope. *Computer Science and Information Systems 11*, 1 (2014), 215–228.

[65] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM 10* (2010), 178–185.

[66] Van Dijck, J. Users like you? theorizing agency in user-generated content. *Media, culture, and society 31*, 1 (2009), 41.

[67] Vanzo, A., Croce, D., and Basili, R. A context-based model for sentiment analysis in twitter. In *Proceedings of COLING* (2014), pp. 2345–2354.

[68] Veeraselvi, S., and Deepa, M. Survey on sentiment analysis and sentiment classification. In *International Journal of Engineering Research and Technology* (2013), vol. 2, ESRSA Publications.

[69] Wang, X. F., and Chen, G. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE 3*, 1 (2003), 6–20.

[70] Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (2010), ACM, pp. 261–270.

[71] Yang, J., and Leskovec, J. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data mining* (2011), ACM, pp. 177–186.

[72] Yang, J., McAuley, J., and Leskovec, J. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE Thirtheenth International Conference on* (2013), IEEE, pp. 1151–1156.

[73] Yu, H.-F., Huang, F.-L., and Lin, C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning 85*, 1-2 (2011), 41–75.

[74] Zhang, H. The optimality of naive bayes. *Association for the Advancement of Artificial Intelligence (AAAI) Vol 1*, No 2 (2004).