Dynamic Analysis of Communication Processes using Twitter Data

Ingo J. Timm,
Jan Ole Berndt, Fabian Lorig

Business Informatics 1
Trier University
54296 Trier, Germany
Email: [itimm,berndt,lorigf]@uni-trier.de

Christof Barth,
Hans-Jürgen Bucher

Media Studies
Trier University
54296 Trier, Germany
Email: [barth,bucher]@uni-trier.de

***Abstract*—Due to the omnipresence of information technology and the increasing popularity of online social networks (OSN), communication behavior has changed. While companies benefit from, i.e., viral marketing campaigns, they are challenged by negative phenomena, like Twitterstorms. Using existing empirical approaches and theories for analyzing the dynamics of social media communication processes and for predicting the success of a campaign is challenging as the circumstances and the access to communication processes have changed. Agent-based social simulation (ABSS) provides approaches to overcome existing restrictions, e.g., privacy settings, and to develop a framework for the dynamic analysis of communication processes, e.g., for evaluating or testing OSN marketing strategies. This requires both a valid simulation model and a set of real world data serving as input for the model. In this paper, a procedure model for the creation of a simulation model is developed and the steps are demonstrated by examples.**

Keywords–Social Network Analysis; Conversation Detection; Networks of Communication; Data Collection and Handling; Simulation Methodology.

I. INTRODUCTION

With the digital revolution initiated by the Internet, social media platforms have gained popularity and have become an inherent part of our private communication. Nowadays, popular OSN, e.g. Facebook, Twitter, or Google+, have more than 1 billion registered users each and the tendency is still rising. Studies report that approximately 28% of the online-time of internet users is spend in OSN [1]. Companies have observed this trend, too, identified the potential of OSN as a platform of aggregated customer contact, and have shifted the focus of many business units to OSN, e.g., customer service or marketing. This has the benefits of facilitating the determination of the customers' demands, of decreasing the efforts of client contact, and of allowing for an identification of trends at an early stage.

A. Dynamics of Communication Processes in OSN

Especially the high degree of connectivity between the users make OSN beneficial for companies, e.g., in terms of word-of-mouth marketing. Compared to the real world, users of OSN are connected with a large average number of people which results in an increased speed of information distribution. This is utilized by marketing strategies of companies to quickly reach a high level of awareness, e.g., in viral marketing campaigns [2]. The self-replicating process of gaining awareness for a certain product or brand is driven by messages or media which are spread by users and which contain information on the entity that is advertised.

However, the effects and mechanisms which are beneficial for companies in terms of viral marketing and for gaining a high level of awareness can also result in harmful consequences. Due to the fast diffusion of information in OSN, negative comments or criticism can be multiplied in an uncontrollable way and cause in a storm of protest. As these storms often occur on Twitter, they are called *Twitterstorms*. A recent example is the *#CrippledAmerica* Twitterstorm. In late 2015, Donald Trump, an American businessman who announced his candidacy for the US presidential election in 2016, mocked a disabled reporter during a political rally while promoting his book "Crippled America". Stuttering stand-up comedian Nina G took this as an opportunity to ask everyone to use the hashtag *#CrippledAmerica* for writing about experiences with disability [3]. As a result, the hashtag's focus shifted from promoting Trump's campaign and book to reports on peoples' experiences with disabilities and negative responses to his statement.

Currently, companies lack methods to direct or end Twitterstorms and thus sometimes inadvertently promote the distribution of negative statements. But the challenge is not only to avoid negative impacts. Also utilizing positive aspects of OSN communication is difficult as traditional concepts of communication can no longer be applied to analyze the dynamics of OSN. The reasons are multilateral communication behaviors as well as an increased number of interpersonal relationships in OSN. Furthermore, the lack of distribution barriers, e.g., (*"death of distance"* [4]), and the increased size of the potential addressees of messages need to be considered.

This does not only challenge companies. Also from a scientific perspective, there is a lack of empirical methods for investigating social mechanisms and dynamics of communication processes as well as for finding explanations in complex systems [5]. Due to their characteristics, compared to traditional offline communication, innovative concepts and techniques are required for analyzing communication processes in OSN [6]. Related research questions arise from the fields of media studies and communication research, as the content and effects of mass media as well as human communication are in focus. Considering standard research methods from these areas, two major challenges can be identified: On the one hand, operators of OSN restrict the access to data and users applying privacy settings to protect their personal data prevent researchers from

 ISBN: 978-1-61208-519-7

accessing relevant information. Thus, field studies can only be conducted when the communication is openly accessible. On the other hand, anonymity and the large number of actors in OSN are factors influencing the behavior of the users. This is why empirical experiments under laboratory conditions are unfeasible, too. It can be assumed that actors will not behave the way they would behave in real OSN, when knowing they are part of an artificial network which is being observed as part of a scientific study. Consequently, alternative approaches are needed for analyzing communication dynamics in OSN, e.g., for evaluating Twitterstorm strategies in advance.

B. ABSS for Analyzing Communication Processes

Computer simulation is a commonly used technique for analyzing complex and inaccessible systems in many disciplines. Here, artificial systems are created by modeling and simulating actors and mechanisms which then can be studied using existing research methods. In contrast to real world systems, simulated systems can be fully accessed, modified, and recreated by the researchers as required. In social sciences, ABSS has been established as a special type of simulation for studying emergent social behavior [7]. By modeling the actors of the real world system as autonomous entities, individual decision-behavior can be simulated and global social phenomena emerge from local interactions of the actors.

For the use in OSN, a data basis as well as a procedure model for the creation of a simulation model are required. The data basis comprises both data about the actors (the users of the OSN), as well as the environment of the actors (the OSN itself). Especially information regarding the types of actors, their actions and goals but also the structure and the opportunities for actions provided by the OSN are needed for creating a suitable simulation model.

Many OSN provide APIs (application programming interfaces) for gathering data about their users and interactions between them. However, APIs provide a large amount of isolated data and the identification of relevant data in terms of ABSS studies is challenging. Thus, the handling of data needs to be assisted and integrated into the process of conducting simulation studies.

This paper presents a first step towards the development of a framework for analyzing communication dynamics in OSN and for testing communication strategies using an ABSS approach. This work particularly focuses on the automated collection, as well as the preparation and selection of relevant communication data from OSN for developing a simulation model as shown in Section IV. In Section V the implementation and evaluation of the approach is described. Here, the syntactical context of the communication will be in focus without further consideration of its semantics. Using the example of Twitter, isolated tweets related to the same topic are selected, individual actors and messages sent by them are derived, and communication dynamics are reconstructed. Furthermore, to evaluate this approach, communication dynamics of Twitterstorms and political discourses are analyzed. Finally, Section VI provides a concluding summary of the findings.

II. FOUNDATIONS

For analyzing the dynamics of OSN communication processes, the act of communication itself but also the structure of OSNs need to be considered.

A. Communication

Human communication can be considered as a sequence of actions of individuals, where the behavior of a sender influences the behavior of a receiver [8]. It can be understood as a process, where the sender uses a set of characters to encode a message, which then is transmitted using an information medium. The receiver uses an own set of characters to decode and interpret the message and returns a feedback using the same mechanism but not necessarily the same medium [9]. However, a message does not necessarily need to be a verbal utterance but can also be nonverbal.

Each message consists of different layers of information. Without further knowledge, a message is only perceived as a set of characters. By adding syntax, the characters become a message, based on rules defining the relationship between characters. The meaning of a message is determined by its semantics. Because of this, the transfer of information can only be achieved if both the sender and receiver share the same semantics. Pragmatics reveal the intention of the message's sender.

The shifting of communication into technical media is accompanied by a loss of information. The transmission of the message is ensured, yet, the receiver does not know whether the message was interpreted correctly. On Twitter, e.g., the platform determines and restricts the communication processes between users and influences the understanding. The result of the communication can only be returned on the same technical way it has been received, by replying to a Tweet using another Tweet. Thus, we can focus on the analysis and simulation of sequences of Tweets and not take nonverbal communication into account at first. For this, it is necessary to know the structure of the network and how communication is made possible. As pragmatics and semantics need to be abstracted for the simulation model, tools for the automated evaluation of messages are needed and are provided by computer linguistics. Even though our example does not focus on the computer linguistic analysis of Tweets, it is an essential part of the model building process as the large amount of data requires an automated approach.

B. Social Networks

In terms of graph theory, the structure of a social network can be described by a set of users (nodes) and relationships between the users (edges), connecting those nodes [10]. Depending on the direction of the relationship, graphs can be unidirectional, defining the direction of the relationship, or bidirectional, connecting two nodes without providing information regarding the direction of the relationship.

For assessing the importance of a node in a graph, e.g., the most influential users of an OSN, centrality measures can be used [11]. The *degree* of centrality corresponds to the total number of edges a node has and can be used as a measure of a node's interconnectedness in a graph. Nodes having a high *degree* (compared to other nodes) are classified as hubs in terms of information diffusion. When considering directed graphs, the *indegree* (number of inbound edges) needs to be distinguished from the *outdegree* (number of outbound edges).

In contrast to this node-specific measure, the *density* is calculated for an entire network or graph. Doing so, it can be used for comparing different graphs. The *density* of a graph is defined by the ratio of the number of existing edges and the

 ISBN: 978-1-61208-519-7

maximum number of edges in case every pair of nodes would be connected by an edge (complete graph).

For simulating communication in OSN, the structure of the network needs to be recreated. A representation of a network using a graph defines the communication channels and the described characteristics give indication of the conditions under which communication is taking place, e.g., who can send messages to whom and how their reach can be assessed.

C. Computational Linguistics

In addition to the structure, OSN consist of messages which are send between the users. For analyzing communication processes, the content of the messages is of relevance, too. It provides the researcher information about the intention as well as the context of communication. Thus, it is desirable to automatically classify the topic of individual messages and communication processes. Doing so, a first impression of the content of communication is given which facilitates the researcher's process of finding and selecting relevant communication processes. Furthermore, a basis for the abstraction of the content for the modeling process is provided. Yet, as the messages consist of natural language, analyzing the content in an automated way is challenging. *Computational linguistics* focuses on the modeling and processing of natural language and provides suitable techniques.

1) Machine Learning: One basic technology used in computational linguistics is *machine learning* which evolved from *artificial intelligence*. In contrast to other algorithms following hard-coded program instructions, *machine learning* algorithms learn from experiences gained from data or from models build from data [12]. Generally, a distinction is made between three types of learning: supervised, unsupervised, and reinforcement learning. While supervised algorithms try to learn rules from example inputs and outputs, unsupervised learning approaches need to find patterns in data on their own. Reinforcement learning takes place in dynamical environments and will not be considered any further in this paper.

2) Content and Lexical Analysis: When using machine learning algorithms for processing natural language, the text first needs to be divided into its linguistic entities. These include words as well as phrases or even entire paragraphs of a text. For separating words, whitespace characters can be used in most segmented writing systems, e.g., those consisting of Latin characters. The entities received when dividing a text are called *n-grams* and are used for creating a model of the language. In this work, n-grams are used for analyzing the mood of messages, i.e., tweets.

For assigning attributes (tags) to words, *part-of-speech tagging* (POST) is applied [13]. Given a text, POST identifies the grammatical categories of each word, e.g., noun, verb, or adjective. This is challenging, as words may appear in different parts of speech at the same time. Yet, analyzing the mostly used nouns, verbs, and adjectives in a large data set, e.g., a set of Tweets, may provide a first impression regarding the most commonly discussed topics.

When analyzing frequencies of words in a text or when indexing documents, a reduction of the words to their base form is needed. *Stemming* aims at reducing words with a similar or identical meaning, but which differ in its suffix, to its word stem. Here, each language requires own stemming algorithms. A commonly used algorithm for the English language is the *porter stemming algorithm* [14].

Summarizing it can be said that for evaluating communication processes in OSN, content and lexical analysis provide information regarding the topic of a conversation and allow for a first assessment of the Tweet.

D. Related Work

The approach presented in this paper is accompanied by related approaches and disciplines where networks of communication and discourses in OSN are analyzed.

Information propagation aims at identifying a group of users which can propagate an unspecified information, i.e., a message, to as many users as possible. Approaches exist where the topics of communication within OSN are explicitly modeled for providing a topic-aware estimation of the propagation probability [15]. Thus, information propagation provides valuable ex-post approaches for analyzing networks of communication but lacks methods for integrating individual and more complex opinion making processes.

Cogan et al. [16] used Twitter data to reconstruct complete conversations around an initial tweet which is given. This enables a more detailed evaluation of conversation topologies, as social interaction models can be compared to OSN. Yet, only isolated and minor conversations lasting up to six hours were analyzed, not larger networks of communication as they occur in Twitterstorms.

For analyzing political discourses among Twitter users, Hsu et al. [17] examined their participation in discussions. The identification of key users was based on the users' public data, e.g., Twitter ID, location, number of tweets, and *follower-followee-networks*, instead of considering the communicative behavior of the users.

Maireder [18] described discourses on Twitter using three perspectives: networking topics, networking media objects, and networking actors. By connecting these perspectives, the author aims at understanding the process of political opinion-making through Twitter using empirical approaches by hand.

These approaches consider the collection and preparation of data as isolated processes for social network analysis. An integration of data handling as a step of an entire research process for generating theories, testing hypotheses or deriving conclusions is not proposed and an adoption of data handling as part of a simulation study is not performed. Therefore, the approach presented in this paper complements existing approaches such that an agent-based simulation of communication processes in OSN is facilitated.

III. Analysis

For analyzing the dynamics of OSN communication processes, a data basis is needed. As the number of existing OSN is large and as OSN differ in structure and mechanisms, the process of data collection differs, too. In this paper, *Twitter* is used as an example platform due to the size of the OSN on the one hand, and the unrestricted access to data on the other hand. Compared to other OSNs like Google+ and Facebook, Twitter's data is not as much affected by privacy settings and can be accessed using the provided API. Still, the communication processes which can be observed on Twitter are of relevance as they affect the general public and have resulted in cross-media phenomena in the past, e.g., the harlem shake [19].

 ISBN: 978-1-61208-519-7

A. Twitter as a Communication Platform

Twitter was founded in 2006 and, compared to other OSN, its unique feature is the limitation of the message ("*tweet*") length to 140 characters. Another difference is how friendships are represented. While most OSN consist of bidirectional relationships between users, meaning two users constitute the *friendship* together, a distinction between *followers* and *followees* is made on Twitter. Here, a user actively and voluntarily decides which other users to *follow* for receiving their status updates in an unidirectional way. Following another Twitter participant makes the following user become a *followee*, yet, the user being followed does not need to follow its *followees*. Thus, a connection between two users does not imply that they exchange information in both directions. In consequence, for analyzing communication dynamics, the directions of the relationships need to be considered.

Besides the user network, the hashtag (#) emphasis Twitter provides is of special interest from a media studies and communication research point of view. When publishing messages, Twitter users can make use of two operators for classifying a message. The #-symbol is used for categorizing messages and for marking keywords of a tweet. This simplifies the researcher's assignment of tweets to a certain topic. Furthermore, Twitter provides mechanisms for replying to other tweets and for addressing a tweet to a certain person. Using the @-symbol followed by the name of a user or by putting the prefix *"RT"* (retweet) at the beginning of a tweet, the identification of dialogs or conversations is supported. Due to these features, Twitter has been widely used for conducting online studies of certain subjects or events, e.g., spread of news [20], the activity of diseases [21] or political communication [22].

B. ABSS of Communication Processes in OSN

For developing a dynamic analysis framework which makes use of simulation techniques, the simulation method needs to be chosen according to the phenomena to be analyzed. A special feature of phenomena occurring in OSN, e.g., Twitterstorms, is that they are emergent [23]. Due to the local interactions of the users on a micro level, global effects occur on a macro level. Yet, they can not (entirely) be explained by the local actions. For analyzing, reproducing, and investigating such emergent phenomena, agent-based computer simulation has been established as a standard means. By modeling real world actors, in this case the users of an OSN, as autonomous software agents, individual behavior and anticipation of behavior on the micro level can be simulated resulting in emergent effects on a macro level [24]. The observation of the global phenomena in combination with the knowledge of the actions and interactions of the actors can then be used for deriving as well as examining scientific explanations regarding the mechanisms of the system. In terms of social sciences, using agent-based actor models for doing social simulation studies is referred to as ABSS [25].

For using ABSS to analyze communication dynamics in OSN, three entities need to be modeled: the users of an OSN (actors), the decisions and actions of the users (behavior), and the connections between the actors (network). While actors and their behavior can be considered as the micro level of the model, the network is a macro phenomenon and can be observed in the real world. Accordingly, an understanding of the macro level needs to be established first, as a basis for further consideration of the actor-based micro level.

During the model-building process, domain expertise is needed for modeling real world mechanisms and processes according to observations or results from discourse and content analysis over time. This information, enriched with theories from software agent technology, can then be technically formalized and used for specifying a multiagent system for simulating OSNs. As a result of this, different artificial scenarios and processes can be observed based on how stochastic events influence the mechanisms. Instead of using the real world system as an object of research, domain-specific research methods can then be applied to the artificial system. Compared to the real world system, a more cost-efficient and restriction-free access to data is provided. Furthermore, variations of the spatial or temporal dimension as well as repetitions of experiments are possible and the real world system is not exposed to any risk or needs not be existent at all. Results of the simulation experiments will be used for refining the model. This enables domain experts to draw conclusions and implications from the model regarding the real world system using specific theories, e.g., for analyzing viral marketing or for preventing Twitterstorms.

The described process results in two interconnected loops of research methodologies where a central interdisciplinary model serves as intermediator. This model is improved and refined stepwise by both disciplines, i.e., simulations and media, until a satisfying state is reached (see Figure 1). The model then can be used in the dynamic analysis framework for simulating OSN and communication processes within them.

IV. Concept

The process of performing a simulation study for analyzing dynamics of communication in OSN can be divided into three major steps (see Figure 3): the acquisition of relevant data, the conduction of the simulation experiments, and the drawing of conclusions from the results of the experiments regarding the real world. In this paper, we focus on the first step, the acquisition of relevant data.

In order to decide which data is relevant for a specific simulation study, the experiments need to be designed in advance. This includes the determination of the methodology of the simulation study as well as the definition of research hypotheses to be tested. After the experimental design has been defined in consultation with the domain experts, e.g., PR experts, relevant data needs to be collected, prepared, and selected accordingly.

A. Data Collection

When gathering OSN data using APIs, most of the data is provided in standardized data formats, e.g., JSON or XML. Due to the structure of the data format, each message or contribution (e.g., tweet or Facebook posting) is transferred as a single piece of information. Additionally, each entity is described by meta data, e.g., a unique ID, the name of the author, a timestamp when it was published, and a reference to which other message it replies.

Twitter provides REST APIs for both, reading and writing data. The access to the API is at no charge and the data can be downloaded as JSON files. Each tweet is characterized by up to 35 attributes, e.g., favorite count and geo coordinates, and up to 500 million tweets are sent per day. Two APIs are intended for the assessment of data, the *streaming* API for accessing the global stream of data and the *search* API

 ISBN: 978-1-61208-519-7

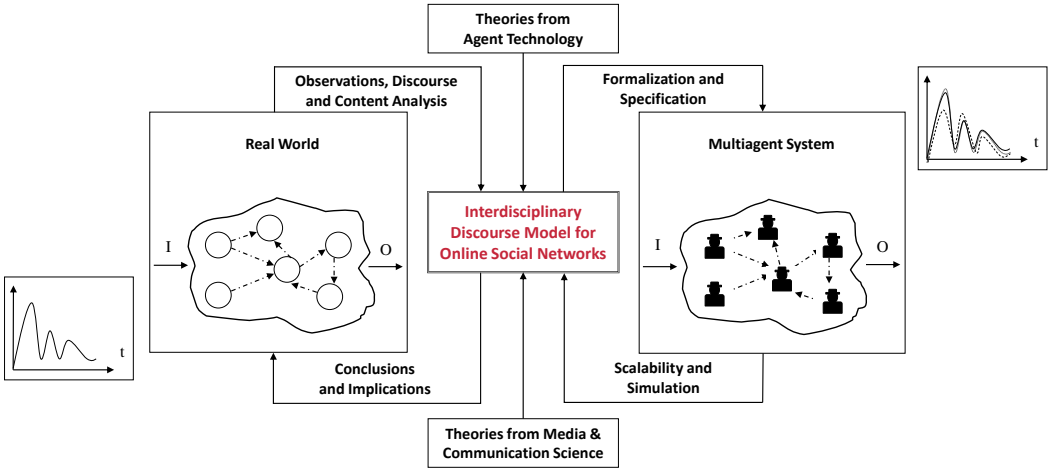

Figure 1. Integrated research method for creating an interdisciplinary model.

allowing queries against a subset of tweets from the past week. Still, both APIs need to be requested with a set of predefined keywords, i.e., hashtags, restricting the results. Here, the tradeoff is the extent of the data. The streaming API provides complete data regarding a hashtag, yet, this results in large datasets which need to be collected and stored in real-time. Technical problems during this process may result in a loss of data, as past data can not be accessed. In contrast to this, the *search* API provides relevant data only which decreases the size of the dataset. The data of the last week can be accessed, which enables a non-real-time collection of data, but the completeness as well as representativity of the provided data are questionable.

Certainly in terms of topics and events that are not discussed using a hashtag which is known in advance, e.g., a Twitterstorm, the advantages and disadvantages of the two APIs are noticeable. The keywords of the real-time streaming API need to be modified in order to capture the tweets of the storm of protest. Yet, when the Twitterstorm is recognized, the beginning has been in the past and thus can not be captured using a real-time API. The search API, in contrast, can be used to collect "popular" tweets of an event which has occurred up to one week ago. Yet, Twitter determines the popularity of a Tweet without providing any information regarding the weighting function being used. Thus, the completeness of the dataset collected using the search API can not be assessed. Consequently, according to the design of experiment, the appropriate API needs to be chosen or a combination of both APIs needs to be used for the collection of data.

B. Data Preparation & Selection

After a dataset has been collected using the API provided by the OSN, it needs to be stored for further processing. In this phase of the data handling, communication processes are identified in the set of isolated tweets, and the content of the communication is analyzed. Furthermore, the network of communication is reconstructed representing related messages and conversations.

1) Conversation Detection & Content Analysis: Topic-related communication processes, i.e., discourses, are considered as coherent dialogs between users or groups of users regarding a certain topic [26]. From a media studies and communication research perspective, the identification and analysis of these discourses within a network of communication is of high relevance. They are the foundation for reconstructing and evaluating topics and opinion-making processes over time.

For discovering discourses in a network of communication, both the conversations between users and the content of the messages need to be analyzed. A conversation is defined by the direction as well as the order of messages which were sent. First, the beginning of a discourse, i.e., the *initial tweet*, needs to be identified. The identification of this tweet in a dataset can be achieved by selecting all tweets, one after another, and checking the following two conditions: 1) Does another tweet exist in the dataset, which is a reply to the selected tweet? and 2) Is the selected tweet no reply to other tweets itself? In case both conditions are fulfilled, a tweet is considered an *initial tweet*. Still, the dataset may contain only a part of a conversation. This might occur, if the initial tweet has not been part of the collection received from the API. In this case, the initial tweet is the one which is a reply itself, yet, the tweet it replies to is not part of the dataset. By iteratively applying this procedure (see Figure 2), communication processes can be identified as shown.

After identifying communication processes between users in networks of communication, an automated analysis of the conversation is desirable due to the large amount of data. Doing so, researchers can get a first impression regarding the type and topic of the conversation. On the one hand, the tonality of tweets can be determined using sentiment analysis, providing information about the mood expressed in the tweets. In terms of discourses, an alternating tonality can be assumed, as two parties talk about the truth of a certain statement. Furthermore, communication processes can be differentiated according to differences of opinion, i.e., pro and contra. On the other hand, an automated analysis of the topic of the conversation can be performed. Content analysis provides techniques for determining commonly used terms in tweets, giving a first impression regarding the potential topic of the conversation.

The tweets are analyzed in two ways. First, the hashtags used in the tweets are identified and collected. An overview of the most commonly used hashtags of a conversation provides

 ISBN: 978-1-61208-519-7

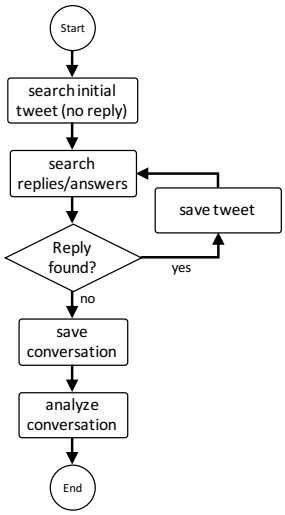

Figure 2. Conversation detection in Twitter dataset.

a first impression regarding the topic of the conversation. As a second step, a POST approach is used for analyzing nouns and adjectives. For doing so, all tweets of the conversation need to be divided into single words. Hashtags can be removed from the set of words, as they have already been evaluated individually and as hashtags often consist of made-up words or abbreviations. Thus, the decision whether a hashtag is a noun or adjective is difficult, too. POST will then be applied to the remaining words to identify nouns and adjectives which occur multiple times. The outcome enables a first assessment of the conversations' topics.

Furthermore, the tonality of a tweet is another indicator for assessing its content. Applying supervised learning algorithms for classifying tweets according to their tonality requires a three-stage approach [27]. As a first step, classification algorithms require a set of training data, which has been classified by hand. Using this data, the learning algorithm is trained and configured for the third step, the automated classification of the remaining tweets. In order to increase the accuracy of the algorithm, a preprocessing of the data should be performed. As the mood of the tweet is assessed by analyzing natural language only, artificial constructs, such as links to websites, @-mentions, and the "RT" prefix can be removed. Doing so, the disturbance of the algorithm can be reduced.

2) Network of Communication: At this point, the dataset contains a large number of individual communication processes. Yet, for analyzing the dynamics of communication, these conversations must not remain separate. Instead, the entire network obtained when merging all individual communication processes is of interest. It contains dependencies between different conversations and provides a chronological order of each conversation. In the following, this topic-specific network of users and messages sent between the users is referred to as *network of communication*.

When reconstructing networks of communication in OSN, the relationships between the users are of relevance. Generally, Twitter provides two kinds of relationships between users: communicative relationships expressed by the use of the *RT* or @ operator and social relationships which are represented by Twitter's *follower-followee-mechanism*.

Analyses of the communicative structure of past Twitterstorms have shown that a small amount of the involved OSN users operate as central nodes and drive the diffusion of the criticism (see Section V). Thus, for reconstructing networks of communication, communicative relationships seem to be most relevant. Social relationships, in contrast, do not contain any information regarding the participation and intensity of communication. Yet, the "communicative power" [26] of a user can be determined by the social interconnectedness of a user. This is relevant when analyzing scenarios that potentially can lead to Twitterstorms, i.e., prospective analysis. In terms of networks of communication, communicative power can be considered as the ability to gain a high level of awareness for a message due to the large number of users the writer is connected with. Accordingly, for reconstructing networks of communication, these types of relationships need to be extracted from the dataset.

Beginning with a large amount of separate tweets and related attributes derived from the Twitter API, a preselection regarding a defined hashtag of interest needs to be performed. At this point, additional filters can be applied for limiting the extent of data, e.g., structure, content or mood filters. Doing so, the dataset is reduced to the relevant tweets directly associated with the topic to be analyzed. Here, the assumption is made that the hashtags mentioned by the tweet imply the topics the tweet is related to, as intended by Twitter. Tweets, that are meant to be related to a topic, yet, do not mention the hashtag in particular, can not be considered as part of the study, as they are not recognized by the API. As a next step, isolated users need to be removed, as they are not part of the network of communication. Accordingly, isolated tweets need to be removed as well, as they are considered not to be of interest to other users. A tweet is classified as *isolated* when it is neither addressing a certain user nor is a retweet or reply to a previous tweet. By considering retweets, circles may occur, as some users tend to retweet their own tweets. These tweets are irrelevant for the network of communication, too.

Based on this cleaned dataset, a directed graph can be generated. In this graph, the nodes represent the users of the OSN and the edges represent the tweets of the users. For simplification purposes, just one type of edges will be used for all three types of communication: retweets, replies, and @-mentions. At that point, the calculation of centrality measures can be performed, e.g., degree or closeness centrality. When visualizing the graph, the researcher gets a first impression of the structure of the network of communication.

As the aim of this process is to create a realistic and valid simulation model, the conceptualization of the simulation model is performed parallel to the data collection and preparation. This facilitates the coordination and enables a harmonization of these two interdependent processes. For one thing, the simulation model is created according to the dataset which has been collected and thus can take account of certain characteristics of the dataset, e.g., involved actor types or specifics of the topic. For another thing, the collection and preparation of data can be adapted to the model ensuring the suitability of the dataset. Starting with a basic conceptualization of the model during the design of experiment and data collection phases, a more detailed conceptualization during the preparation and selection phase is done. This results in the creation of an applicable simulation model which matches the acquired data as it has been developed based on them.

 ISBN: 978-1-61208-519-7

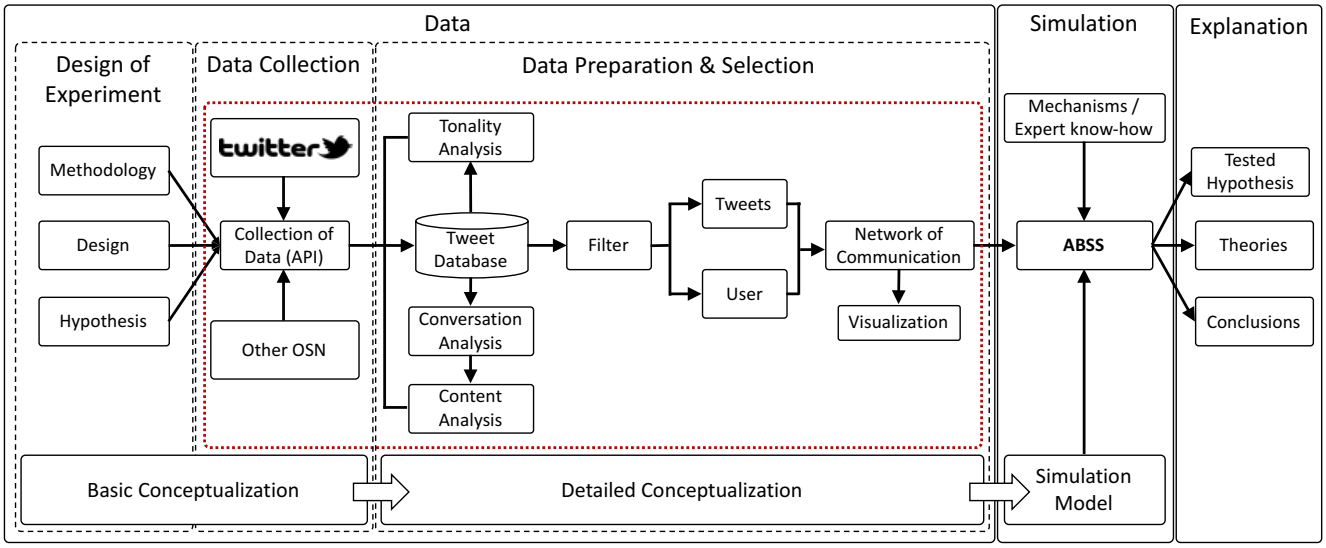

Figure 3. Procedure model for collecting, editing, and aggregating OSN data for ABSS studies.

The data collected and prepared in the previous steps can now serve as input for the simulation model that has been developed simultaneously. At this step, ABSS experiments can be conducted using the results of the previous process step. In addition, expert know-how is needed in order to validate and verify the simulation model as well as to interpret the results of the experiments. This includes proving or disproving of the hypotheses defined during the design of experiment phase as well as deriving conclusions or theories from the results.

V. Implementation and Evaluation

As a proof of concept and for evaluating the procedure model proposed in Section IV, the process of collecting and preparing data for ABSS studies is implemented. Furthermore, the feasibility of the implementation is evaluated by analyzing the datasets of two Twitterstorms.

A. Implementation of the Framework

For querying the Twitter API, a *PHP* script has been developed and used. The results are formatted as *JSON* objects and include all necessary information regarding the tweet itself as well as the user which has been the author of the tweet. The data is stored in a *MySQL* database which is used for the central data management.

For the preparation of the data, existing software packages can be used providing basic algorithms, e.g., machine learning or part-of-speech tagging algorithms. A number of frameworks exist, e.g., Apache Mahout or Scikit-Learn. However, due to the programming language it is implemented with and the large amount of preimplemented algorithms, the *DatumBox* framework [28] has been chosen for this implementation. *DatumBox* is a framework which provides natural language processing and classifying services written in *JAVA*. It focuses on social media monitoring as well as text analysis and quality evaluation in online communities. The learning algorithms of the *DatumBox machine learning framework* have been used for this implementation, as the framework can handle large datasets and is open-source. The implementation of the *support vector machine* uses *LIBSVM* [29], a widely used open-source implementation of SVM. Furthermore, *Apache Lucene* [30] is used as text search engine, which is open-source and used by large companies, e.g., Twitter, for real-time search.

After collecting raw communication data, this implementation allows for performing tonality analyses using the algorithms of the *DatumBox* framework and *Apache Lucene*. In order to obtain the required training data, a number of tweets needs to be classified by human beings, after they have been edited. This training data as well as SVM, n-gram, and stemming algorithms provide a classification of the tweets regarding their mood.

The conversation detection has been implemented as shown in Figure 2, followed by an analysis of the conversations' topics. The results of both analyses are then saved in the central database, too.

As a next step, for reconstructing networks of communication, the tweets of the database are filtered regarding the hashtags of interest. Additionally, the involved users are loaded from the database and a graph is created. The users serve as nodes, while each tweet is illustrated as a directed edge indicating the direction of the communication. For a reply, the edge would point from the user who replied to it to the author of the original tweet.

B. Analysis of the #pegida Twitterstorm

For evaluating the proposed approach, Twitter data has been collected since the beginning of 2015. For doing so, the hashtags of current topics of online news media have been used as keywords. During this period, 18 Mio. tweets containing 8 Twitterstorms have been recorded. Both *#pegida* and *#deflategate* are hashtags of considerable communication processes which took place on Twitter during this period of time.

The evaluation of the conversation analysis requires a highly discursive topic, providing conversations with a high depth. For this reason, the social media echo of the *Pegida* protests has been chosen as dataset containing 3.2 Mio. Tweets [31]. Pegida is a right wing political movement that was founded in Dresden, Germany in October 2014 and opposes

 ISBN: 978-1-61208-519-7

the perceived "Islamisation" of the Western world. Hence, due to the formation of opposing interest groups supporting or rejecting Pegida's point of view, opinions are divided and the formation of discourses is facilitated.

Analyzing the dataset, 19 685 conversations were identified consisting of nearly 51 000 tweets. Conversations can be classified by the number of replies as well as by the depth (steps) of the conversation. Figure 4 shows the distribution of the conversations by number of replies and depth. Conversations of a depth higher than 10, meaning that two users wrote 5 messages each replying to the previous message of the other user, are not existing whereas 136 conversations have more than 10 replies.

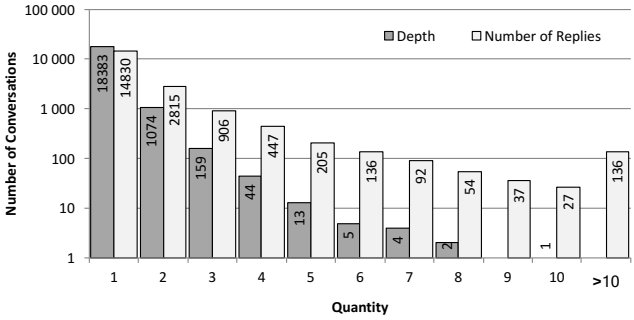

Figure 4. Distribution of conversations by number of replies and depth of conversation frm the #pediga analysis.

The structure of conversation trees can be divided into two major groups: *paths* and *stars* [32]. A *star* is defined by a low depth of the tree combined with a high number of tweets, consequently, a high amount of replies to one or a few tweets. In contrast, *paths* have a high depth while the total number of tweets is low.

Further analysis of the data showed that two types of *stars* exist in the dataset that differ in the number of involved users. The most extensive conversations of the dataset, consisting of 107 and 100 tweets, are the result of only 3 resp. 2 users. On closer examination, these conversations were classified as spam. Thus, we assume that for conversations on Twitter the ratio between the number of tweets and the number of involved users can serve as a indicator for spam. This assumption was strengthened by a manual analysis of the dataset. For most spam conversations, the ratio between users and tweets was at least 1 to 10. Accordingly, this type of star can be referred to as *spam star* and is not relevant for further analysis.

A second type of stars exists where the ratio is inverted. In this case, the number of tweets and the number of users is almost equal implying that the majority of users commented on the conversation only once. In over 90% of the conversations which were detected in the dataset, 90% of the tweets have been written by different users. Consequently, most of the tweets have not been replied to. Even though these stars represent relevant conversations, they may not be considered as discourses, as the *"back-and-forth"* character of discourses is missing.

The paths, in contrast, are what we consider to be discursive behavior. Two or more users respond to each other's Tweets and constitute a conversation. By merging both stars and paths, the network of communication can be reconstructed for further analysis. Furthermore, for the modeling of agent behavior, it appears that the communication rather than the exchange of opinions is in focus. This is triggered by an initial tweet and results in a *Fire-and-forget* behavior of the users.

C. Analysis of the #deflategate Twitterstorm

The reconstruction of the networks of communication is evaluated using the dataset of the *#deflategate* Twitterstorm. Due to the limited timespan of a Twitterstorm, the collection of a complete dataset is simplified. Furthermore, analyzing a Twitterstorm's network of communication is of interest, as central users or tweets can be identified.

The *#deflategate* storm started three days after the 49th NFL Super Bowl and was triggered by a Tweet of the journalist Chris Mortensen, claiming 11 of the 12 footballs were underinflated [33]. As each team plays with separate footballs and as the hosting team supplies the balls, this appeared to have happened on purpose, to influence the behavior of the ball when thrown, kicked or caught. 17 621 tweets from 9 870 users have been collected during the *#deflategate* storm. Out of this, 41 tweets reply to themselves and 4 577 users are isolated and thus were removed. Consequently, the network of communication consists of 5 293 users and 6 067 tweets.

Two central nodes can be identified in the network of communication. This observation is confirmed when comparing the *degree* centrality of the nodes. While the average *degree* is 1, a user named *TomBradysEgo* (Twitter User-ID: 317170443) is having the maximum *degree* of 509. *TomBradysEgo* is a parody account on Tom Brady, the quarterback of the New England Patriots, having 235 000 followers and posting an average of 113 tweets per month. During the Twitterstorm, 39 tweets were published by the account. Due to the high *outdegree*, 97.4% of the total *degree* of the node, in combination with the low number of published tweets, it can be assumed that the user's tweets have often been retweeted. Thus, a central role of *TomBradysEgo* can be implied and the account can be classified as a hub.

Similarly, the user named *brownjenjen* (Twitter User-ID: 2453787236) has a *degree* of 485 and is an American blogger. Having only 23 000 followers, *brownjenjen* published 43 tweets during the Twitterstorm. Due to the *outdegree* of 100%, a large number of retweets can be assumed, too. As the account does not reply to other tweets and participates in different topics, it can be classified as a hub, too.

The important role the two accounts play for the Twitterstorm clarifies, when removing the two nodes and the related communication from the network of communication. Doing so, the density of the graph is reduced by 12.46% which can be compared to a reduction of the communication by the same extent. The union of the *ego-centered networks* of the two central nodes illustrates their maximal neighborhood, i.e., all nodes that can be reached from the central nodes. Here, 69.69% of the communication of the storm is linked to the two central nodes, showing their overall impact. According to this, a more detailed consideration of these two users seems promising in terms of social network analysis.

For both topics, *#pegida* and *#deflategate*, the feasibility of the approach proposed in this paper has been shown. In terms of content and discourse analysis as well as reconstruction of networks of communication, preliminary results assisting the selection of relevant data for subsequent studies were generated. Thus, when simulating emergent OSN phenomena, the different reach of agents needs to be considered. Some

 ISBN: 978-1-61208-519-7

agents need to serve as hubs for pushing the diffusion of messages.

VI. Conclusions and Future Work

In this paper, a first step towards the development of a dynamic analysis framework for OSN communication processes is proposed. A major challenge is the collection as well as the preparation and selection of relevant data, which is addressed by the presented approach. Currently, the analysis of a set of collected data for interesting phenomena for further consideration is done by hand. Our concept aims at providing assistance functionalities, by automating the handling of data for the preparation of simulation studies.

For gaining a first overview of the dataset of isolated messages, conversations between users are detected, the content and tonality of the messages are assessed, and the network of communication is reconstructed. Using the examples of *#deflategate* and *#pegida*, the process of data collection as well as data preparation and selection has been implemented and evaluated. The network of communication has been visualized and central nodes of the communication graph have been identified automatically.

This work is only a first step towards a framework for analyzing the communication dynamics of OSN. Besides the aspect of data collection and preparation, which has been subject of this paper, the creation of the simulation model as well as the integration of data and model for conducting simulation experiments need to be considered. For building an agent-based simulation model, actor types need to be derived from social network data, too, and a consideration of communication across different networks is desirable. Furthermore, from a media studies and communication research perspective, a more detailed specification of the interactions between users and the subject of communication are needed for conducting sound ABSS studies.

Acknowledgments

We would like to acknowledge our master students Nils Dammenhayn, Stephanie Rodermund, Christopher Schulz and Nicolas Schulz for contributing to this work.

References

[1] J. Mander, "Daily time spent on social networks rises to 1.72 hours," https://www.globalwebindex.net/blog/daily-time-spent-on-social-networks-rises-to-1-72-hours, [retrieved: 09/16].

[2] J. Kirby, Connected marketing: the viral, buzz and word of mouth revolution. Amsterdam: Butterworth-Heinemann, 2010.

[3] Nina G, "Disability Community Tweet-in," https://ninagcomedian.wordpress.com/2015/12/01/donald-trump-tweet-in-for-crippledamerica/, [retrieved: 09/16].

[4] E. Tranos and P. Nijkamp, "The death of distance revisited: Cyber-place, cyber-place, physical and relational proximities," Journal of Regional Science, vol. 53, no. 5, Dec. 2013, pp. 855–873.

[5] R. Mayntz, "Mechanisms in the analysis of social macro-phenomena," Philosophy of the social sciences, vol. 34, no. 2, 2004, pp. 237–259.

[6] F. Lorig and I. J. Timm, "How to model the human factor for agent-based simulation in social media analysis?" in Proceedings of the 2014 ADS Symposium (part of SpringSim multiconference). SCS, 2014, p. 12.

[7] D. Helbing, Social self-organization: Agent-based simulations and experiments to study emergent social behavior. Springer, 2012.

[8] C. R. Berger, "Interpersonal communication," The International Encyclopedia of Communication, 2008.

[9] C. E. Shannon, "A mathematical theory of communication," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, 2001, pp. 3–55.

[10] F. Vega-Redondo, Complex Social Networks. Cambridge University Press Cambridge, MA, 2007.

[11] L. C. Freeman, "Centrality in social networks conceptual clarification," Social networks, vol. 1, no. 3, 1978, pp. 215–239.

[12] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, no. 2, 1988, pp. 95–99.

[13] A. Voutilainen, "Part-of-speech tagging," The Oxford handbook of computational linguistics, 2003, pp. 219–232.

[14] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, 1980, pp. 130–137.

[15] C. Zhang, J. Sun, and K. Wang, "Information propagation in microblog networks," in Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. ACM, 2013, pp. 190–196.

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

[17] C.-l. Hsu, S. J. Park, and H. W. Park, "Political Discourse Among KEy Twitter Users: The Case Of Sejong City In South Korea," Journal of Contemporary Eastern Asia, vol. 12, no. 1, 2013, pp. 65–79.

[18] A. Maireder, "Political Discourses on Twitter: Networking Topics, Objects and People," in Twitter and Society. Peter Lang, 2013.

[19] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 450–453.

[20] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks." ICWSM, vol. 10, 2010, pp. 90–97.

[21] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," PloS one, vol. 6, no. 5, 2011, p. e19467.

[22] A. Maireder and S. Schlögl, "24 hours of an# outcry: The networked publics of a socio-political debate," European Journal of Communication, 2014, pp. 1–16.

[23] J. Goldstein, "Emergence as a construct: History and issues," Emergence, vol. 1, no. 1, 1999, pp. 49–72.

[24] J. O. Berndt and O. Herzog, "Anticipatory behavior of software agents in self-organizing negotiations," in Anticipation Across Disciplines. Springer, 2016, pp. 231–253.

[25] P. Davidsson, "Agent based social simulation: A computer science view," Journal of Artificial Societies and Social Simulation, vol. 5, no. 1, 2002.

[26] J. Habermas, Between facts and norms: contributions to a discourse theory of law and democracy, ser. Studies in contemporary German social thought. Cambridge, Mass: MIT Press, 1996.

[27] S. Abney, Semisupervised learning for computational linguistics. CRC Press, 2007.

[28] DatumBox Framework, http://www.datumbox.com, [retrieved: 09/16].

[29] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm, [retrieved: 09/16].

[30] Apache Lucene, http://lucene.apache.org/core, [retrieved: 09/16].

[31] E. Crecsi, "How Germans documented Pegida's far-right protests on social media," http://www.theguardian.com/world/2015/jan/06/how-germans-documented-pegidas-far-right-protests-on-social-media, [retrieved: 09/16].

[32] P. Cogan, M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci, "Reconstruction and analysis of twitter conversation graphs," in Proceedings of the HotSocial ACM International Workshop. ACM, 2012, pp. 25–31.

[33] A. Jaafari, "The goodell, the bad, and the ugly: The minimal level of integrity in the NFL's disciplinarian of players," Social Science Research Network (SSRN), 2016.

 ISBN: 978-1-61208-519-7