PROBABILISTIC RELATIONAL MODELS FOR SENTIMENT ANALYSIS IN SOCIAL NETWORKS

FEDERICO ALBERTO POZZI

DOCTOR of PHILOSOPHY

Department of Informatics, System and Communication
Faculty of Maths, Physics and Natural Science
University of Milano-Bicocca

October 2014

Advisor: Prof. Enza Messina
Co-Advisor: Dr. Elisabetta Fersini

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

ABSTRACT

The huge amount of textual data on the Web has grown in the last few years rapidly creating unique contents of massive dimensions that constitutes fertile ground for Sentiment Analysis. In particular, social networks represents an emerging challenging sector where the natural language expressions of people can be easily reported through short but meaningful text messages. This unprecedented contents of huge dimensions need to be efficiently and effectively analyzed to create actionable knowledge for decision making processes. A key information that can be grasped from social environments relates to the polarity of text messages, i. e. the sentiment (positive, negative or neutral) that the messages convey. However, most of the works regarding polarity classification usually consider text as unique information to infer sentiment, do not taking into account that social networks are actually networked environments. A representation of real world data where instances are considered as homogeneous, independent and identically distributed (i.i.d.) leads us to a substantial loss of information and to the introduction of a statistical bias. For this reason, the combination of content and relationships is a core task of the recent literature on Sentiment Analysis, where friendships are usually investigated to model the principle of homophily (a contact among similar people occurs at a higher rate than among dissimilar people). However, paired with the assumption of homophily, constructuralism explains how social relationships evolve via dynamic and continuous interactions as the knowledge and behavior that two actors share increase. Considering the similarity among users on the basis of constructuralism appears to be a much more powerful force than interpersonal influence within the friendship network. As first contribution, this Ph.D. thesis proposes Approval Network as a novel graph representation to jointly model homophily and constructuralism, which is intended to better represent the contagion on social networks. Starting from the classical state-of-the-art methodologies where only text is used to infer the polarity of social networks messages, this thesis presents novel Probabilistic Relational Models on user, document and aspect-level which integrate the structural information to improve classification performance. The integration is particularly useful when textual features do not provide sufficient or explicit information to infer sentiment (e. g., I agree!). The experimental investigations reveal that incorporating network information through approval relations can lead to statistically significant improvements over the performance of complex learning approaches based only on textual features.

PUBLICATIONS

CONFERENCES

Pozzi, F. A., Maccagnola, D., Fersini, E., and Messina, E. (2013). Enhance user-level sentiment analysis on microblogs with approval relations. In *AI*IA 2013: Advances in Artificial Intelligence*, volume 8249 of *Lecture Notes in Computer Science*, pages 133–144. Springer International Publishing.

Pozzi, F. A., Fersini, E., and Messina, E. (2013). Bayesian model averaging and model selection for polarity classification. In *Proc. of the 18th International Conference on Application of Natural Language to Information Systems*, LNCS, pages 189–200.

Fersini, E., Messina, E. and **Pozzi, F. A.** (2014). Subjectivity, Polarity And Irony Detection: A Multi-Layer Approach. *CLIC-it 2014*.

JOURNALS

Fersini, E., Messina, E. and **Pozzi, F. A.** (2014). Sentiment Analysis: Bayesian Ensemble Learning. *Decision Support System*. In press.

Fersini, E., Messina, E. and **Pozzi, F. A.** (2014). Expressive signals in social media languages to improve polarity detection. *Information Processing & Management, IP&M 2014. Special Issue on Sentiment Analysis*. Under revision.

WORKSHOPS

Pozzi, F. A., Fersini, E., Messina, E., and Blanc, D. (2013). Enhance polarity classification on social media through sentiment-based feature expansion. In *WOA@AI*IA*, pages 78–84.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

NLP Natural Language Processing

SA Sentiment Analysis

SM-SA Social Media Sentiment Analysis

SRL Statistical Relational Learning

PRMs Probabilistic Relational Models

N-DAG Normalized Directed Approval Graph

A-DAG Augmented Directed Approval Graph

H-DAG Heterogeneous Directed Approval Graph

MV Majority Voting

BMA Bayesian Model Averaging

DIC Dictionary-based classifier

NB Naïve Bayes

ME Maximum Entropy

SVM Support Vector Machines

CRF Conditional Random Fields

S^2-LAN Semi-supervised Sentiment Learning by Approval Network

EM Expectation-Maximization

LDA Latent Dirichlet Allocation

LSA Latent Semantic Allocation

PLSA Probabilistic Latent Semantic Analysis

ASUM Aspect and Sentiment Unification Model

JST Joint Sentiment/Topic model

TSM Topic Sentiment Mixture model

NAS Networked Aspect-Sentiment model

1 INTRODUCTION

"Those who can save the contents of the tweets do not own only a business model, but also a new form of power."
— C. Frediani, Journalist

The great diffusion of social media and their role in the modern society, represent one of the more interesting novelty in these last years, capturing the interest of researchers, journalists, companies and governments. The dense interconnection that often arises among active users generates a discussion space that is able to motivate and involve individuals of a larger Agora, linking people with common objectives and facilitating diverse forms of collective action. This gives rise to what is called "*individualism on the net*": instead of counting always on a single reference community, thanks to the social media becomes possible to move among more people and resources, often heterogeneous. Therefore it appears that social media are creating a digital revolution. The most interesting aspect of this change is not solely related to the possibility of promoting political participation and activism. The real social revolution invests the lives of every single individual. It is the freedom to express ourself, to have our own space in which to be ourself, or what you would like to be, with few limits and barriers. The social media revolution is then able to talk about our emotions and opinions not only to ourself, but especially to those around us, interacting with them, opening a window on others' respective worlds, and snooping into their lives.

Paradoxically, all this happens while we live in a society where it is even more difficult to know the names of our neighbors, and where the right to privacy becomes an imperative to which submit to, to subsequently communicate everything (strictly on-line) to the whole world: love, unforgettable days and daily failures. Because on social media we effectively end up telling the whole (or most) of our life: happiness for the birth of a child, the anger for a train delay, the pre-Christmas shopping or the choice made in the secrecy of the voting booth. It is then not surprising if researchers started to discuss the methods by which to capture this vast sea of information. Because the data on the net, if properly collected and analyzed, allows not only to understand and explain many complex social phenomena, but also to predict them.

The growing interest of Sentiment Analysis is related to the possibility of exploiting its results in several tasks

Relationships provide useful information to enhance traditional learning approaches

One of the most related task is called Sentiment Analysis (SA), where "What people think" is captured and analyzed by complex models. In the past, people asked their friends for advice and opinions related to any kind of topic, such as which restaurant is the best one in the city, or who would they vote for in the next elections. For this reason, instead of looking at Text Mining from the traditional point of view which consider the text meaning, in this thesis the emotional view is taken into account: the sentiment. Precisely, the aim of SA is to define automatic tools able to extract subjective information, such as opinions, sentiments, evaluations or attitudes, from texts in natural language. Extracting automatically subjective information could help individuals, companies and organizations to make more informed decisions. The growing interest of SA is related to the possibility of exploiting its results in different tasks, such as: understand and forecast the sentiment of financial markets (Mitra and Mitra, 2011), manage business intelligence tasks related to users feedback (Pang and Lee, 2008) or sound out public opinion during political campaigns (O'Connor et al., 2010). However, today the situation is changed through the web: we are able to access opinions and experiences of a greater pool of people, which could be neither our personal acquaintances nor well-known professional critics. Blogs, microblogs and online social networks are constantly flooded with opinions about a multitude of topics such as politics, sports and other "buzz" topics that pop up daily on news media. This kind of data comes from different sources such as news, editorials and, more generally, user generated content. Typical examples of the latter are social network contents, which are growing ever vaster and are gaining ever more popularity. Considering the evolution of the sources where opinions are issued, the strategies available in the current state of the art are therefore no longer effective for mining opinions in this new and challenging environment. A contribution that could allow us to overcome the limitation of the available state-of-the-art approaches comes from the Statistical Relational Learning (SRL) research field. Recent advances have convincingly demonstrated that relationships provide useful information to enhance traditional learning approaches (Deng et al., 2013; Tan et al., 2011).

1.1 THESIS OUTLINE AND CONTRIBUTIONS

The thesis is organized as follows. In Chapter 2, a comprehensive survey on social networks is presented. The difference between Social Media and Social Network is first discussed, to subsequently present the characteristics, differences and statistics of the most popular social networks. In Chapter 3, the literature review regarding traditional Sentiment Analysis, Sentiment Analysis on Social Media and their applications, polarity classification and joint aspect-sentiment extraction techniques are presented and discussed in detail. In Chapter 4, the Approval Network able to model similarities among users is proposed and two sociological processes underlying the network modeling are investigated. Moreover, the existing types of relational data representation, the graph transformation techniques and the Probabilistic Relational Models able to deal with relational data are presented. Three main contributions for polarity classification are presented and detailed experimental investigations on several benchmark datasets are discussed. The first contribution of this thesis is related to the definition and modeling of Approval Network (Chapter 4). It is constructed using approval relations among users. The general idea behind Approval Network is that a user who approves (e. g., by 'likes' on Facebook or 'retweets' on Twitter) a given message is likely to hold the same opinion of the author. Thus, "approving" usually means agreeing with the original user: the more are the approvals between two users upon a particular topic of interest, the higher is their agreement on that topic.

As second contribution, an ensemble of different classifiers for polarity classification at document-level has been proposed (Chapter 5) based on the bayesian paradigm. The idea behind Bayesian Model Averaging (BMA) is to exploit the characteristics of several independent learners by combining them in order to achieve better performance than the best baseline classifier. It has been demonstrate that BMA leads to more robust and accurate classification.

As third contribution, a semi-supervised sentiment learning approach has been introduced to deal with polarity classification at document and user-level (Chapter 6). Semi-supervised Sentiment Learning by Approval Network (S^2-LAN) combines text and Approval Network: given a small proportion of users already labeled in terms of polarity, it predicts the sentiments of the remaining unlabeled users by combining textual information and Approval Network directly in the probabilistic model.

As fourth and last contribution, an unsupervised probabilistic model called Networked Aspect-Sentiment model (NAS) is proposed to simultaneously extract aspects and classify sentiments from textual messages (Chapter 7). It incorporates approval relations to perform the sentiment classification and aspect extraction tasks simultaneously.

Finally, in Chapter 8 conclusions are derived and some future works are highlighted.

2 SOCIAL NETWORKS

> *"Relationships are all there is. Everything in the universe only exists because it is in relationship to everything else. Nothing exists in isolation. We have to stop pretending we are individuals that can go it alone."*
> — M. Wheatley, Writer

Online social networks such as *Twitter*, *Facebook* and *LinkedIn* have become very popular in recent years, mainly because of the increasing proliferation of devices such as personal computers, smart-phones and tablets. Online social networks have lead to a big explosion of network-centric data in a wide array of scenarios.

In general, a social network is defined as a network of interactions (or relationships), where the nodes consist of actors/entities, and the edges consist of the relationships between these actors/entities. Nowadays, the concept of social network is commonly restricted to the specific case of an internet-based social network (e. g., Facebook). However, the problem of social networking has been extensively studied often in the field of sociology. The studies on social network analysis have historically preceded the advent and popularity of computers and the internet. A classic example of this is the study of Milgram (Milgram, 1967) in the sixties, who hypothesized the likelihood that any pair of actors on the planet are separated by at most six degrees of separation (i. e. one arbitrary actor is connected to any other by at most six connections). Such hypotheses have largely remained conjectures over the last few decades (Aggarwal, 2011). However, the development of online social networks provided an opportunity to test such hypotheses at least in an online setting (i. e. the "small world phenomenon"), showing that the average path length between two MSN messenger users is 6.6 (Leskovec and Horvitz, 2008).

A social network is defined as a network of interactions

The data-centric impetus has lead to a significant amount of research: it allowed researchers to point both the statistical and computational focus in analyzing large amounts of online social network data. In the following, the different settings for social network analysis are enumerated:

- The most classical definition widely considered in the field of sociology, reports that a social network is based purely on human interactions. These studies have traditionally been conducted

A social network is based purely on human interactions with painstaking and laborious methods for measuring interactions between entities by collecting the actual data about human interactions manually. An example is the six-degrees-of-separation experiment by Milgram (Milgram, 1967), who used postal mail among participants through the use of locally chosen forwards of the mail. Such experiments are often hard to conduct in a completely satisfactory way, because the actors may have response rates which cannot be cleanly modeled in terms of social interaction behavior. For example, the Milgram experiment results have often been questioned (Kleinberg, 2006) because of the low forward rate of the letters which never reached the target. The social analysis of such networks has also been modeled in the field of cognitive science, where the cognitive aspects of such interactions are utilized for analytic purposes. Much of the research in the traditional field of social networks has been conducted from this perspective. A number of books (Carrington et al., 2005; Wasserman and Faust, 1994; Watts, 2004) provide an understanding of this perspective.

- Telecommunications, electronic mail, and electronic chat messengers (such as Skype or Google Talk), can be considered an indirect form of social networks, because they are naturally modeled as communications among different actors. This data can be used for extensive analysis of such social networks.

- In recent years, a number of websites have explicitly arisen to model the interactions among different actors. Some examples are Facebook, MySpace, and LinkedIn. In addition, websites which are used for sharing online media content, such as Flickr and YouTube, can also be considered indirect forms of social networks. We note that such social networks are extremely rich in terms of the amount of content such as text, images, audio or video. Such content can be leveraged for a wide variety of purposes.

- Finally, a number of social networks can also be constructed from specific kinds of interactions in different fields. An example is the bibliographic network that can be constructed from either co-authorship or citation data among scientific papers. Interesting results could be achieved through the combination of connections and the content of the publications.

2.1 SOCIAL MEDIA OR SOCIAL NETWORK?

To understand what meaning is adopted for social media, to differentiate among different types of social media and to identify the particular sub-class of social media which this thesis is focused on, it is useful to begin our discussion by introducing the concept of "social networks". By social network we refer to any structure, formal or informal, consisting of a group of people or organizations, together with their respective relationships (Scott, 2000). Usually, a graphical representation of a social network is given by "*nodes*", corresponding to the actors who operate in that network, along with the *connections* among these nodes, which may be more or less dense depending on the intensity of social relations existing among them. These reports can be either *explicit*, as in the case of classmates or family ties, and *implicit*, as with friendships, and may originate and carry out off-line (i. e. in the real world) and on-line (i. e. in the network).

A social network is given by nodes, corresponding to the actors who operate in that network, along with the connections among these nodes

Social media are virtual platforms that allow you to create, publish and share content, which are generated directly by their users (Yu and Kak, 2012). In this sense, social media differ from traditional media, such as newspapers, books and television because of their horizontality with respect to the possibility (and faculty) to publish content. If, for example, a newspaper generally contains only the news written by its reporters, in the case of social media the barrier of entry to the "production" of a text is virtually absent: just a computer (or mobile phone) with free internet to do so. There are various types of social media (Kaplan and Haenlein, 2010). For example, *Wikipedia*[1], a source of global information of large influence, represents a particular type of social media that goes under the name of "*collaborative project*". In more details, collaborative projects directly affect users who are called to work together with the aim to produce content that will then become accessible to the whole network. But Wikipedia is not a social network, which must meet three minimal conditions: (1) there must be specific users of the media, (2) these users must be connected each other, and (3) there must be the possibility of interactive communications among them. Other types of social media are represented by the "*content communities*", i. e. platforms where users can share specific content with other members of the online community as occurs for videos on *YouTube*[2]. Moreover, a blog in which the writers recount the daily events of their lives or express their own ideas unilaterally (i. e. they do not satisfy any condition of "interactivity"), cannot be defined technically as a social network.

In this thesis, only those social media which have also a form of social network will be treated. For this reason, when the wording social media is used, this subset is taken into account. These types of social media play two basic functions: on one side they produce relationships, on the other side produce content. With regard to the relationships, these can reflect already existing or completely new (based on shared interests or common activities) social networks. Regarding content, they can be created from scratch by the user, but

1 en.wikipedia.org
2 www.youtube.com

also shared and/or exchanged. *Facebook* and *Twitter* are currently the most popular social media.

2.1.1 *Facebook, Twitter and Google+: characteristics and differences*

As regards the admitted social interactions, a **Facebook**[3] user must become "friend" of another user (and be accepted as such by the latter, i. e. the friendship is bilateral) in order to access to the information published by this second user and to be able to write, for example, messages on its timeline. An exception are the profiles or features that users decide to make "public". In this case, the information published become freely accessible to all Facebook users (as well as to those who have not an account). Finally, a Facebook user can express his/her pleasure or interest in the activities of other users, initiatives, campaigns, brands or business/institutional accounts through the '*like*' function (Joinson, 2008).

Another very popular social network is **Twitter**[4]. It was born two years later, in 2006. Unlike Facebook, each user on Twitter can only share short text messages up to a maximum of 140 characters, called '*tweets*'. These are updates that are shown in the user's profile page and are automatically forwarded to all those who have registered to receive them (i. e. the followers of the user). Thanks to the constant production of ideas and textual content, which may include also images, links, and short videos, Twitter is considered as a social network that generates microblogging. Unlike Facebook, the "*Tacit consent*" rule is applied on Twitter: a user can "follow" another without his/her consensus, unless he/her has not decided to make a private account. If users deem it appropriate have the ability to "lock" their followers, making it impossible to continue to follow them. In addition, a user can freely direct a public message to any other Twitter user, regardless of their relationship. This can be done through the simple addition in the message of the sign "*at*" (@) succeeded by the name corresponding to the account of the user to whom you want to target your tweet. In this case, the user whose account is included in the tweet will receive an automatic notification with the accompanying text. On Twitter there is also the possibility to spread the entire tweet of another user by sending tweets to all your followers (the so-called "*retweet*"). Moreover, messages posted on Twitter can be labeled with the use of one or more 'hashtag', i. e. words or combinations of words preceded by the pound sign (#). By labeling a message with a hashtag, a hyperlink to all the other messages which contain the same hashtag is automatically created. A Twitter user can thus easily access to these posts, regardless of the fact that they come from the users who you are following or not. Despite the Twitter user base

Twitter is considered as a social network that generates microblogging

3 www.facebook.com
4 www.twitter.com

is only a fraction of that of Facebook, this social media, because of its openness and horizontality, is becoming a source of real-time news extremely influential both for the entire network and the traditional media.

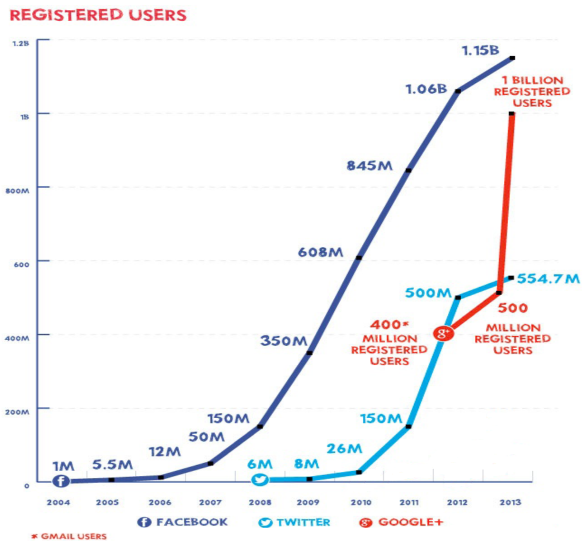
Finally, **Google+** is the youngest of the three, having been launched by Google Inc. in June 2011. Compared to other social media, Google+ includes new multimedia content, such as the ability to initiate audio and video sessions through the 'hangouts', virtual rooms where it is possible to communicate with multiple users at the same time. The system of contacts, equivalent to friends on Facebook or followers on Twitter is organized into 'circles' freely creatable and modifiable by the user. The circles can be labeled as 'acquaintances', 'work' and so on. In order to add a new connection to the user's circles, the permission of the user is not requested (as Twitter), but each user has the ability to customize the information shared with the various connections according to the circles settings. The model of Google+ in this sense is halfway between Facebook and Twitter.

Figure 1 shows that Facebook appears to be the most used platform in terms of total users with over 1 billion active users, and with a penetration rate[5] of 51%, followed by Google+ (1 billion registered users, 500 million active users, penetration rate of 26%), and Twitter (550 million registered users, 297 million active users, penetration rate of 22%).

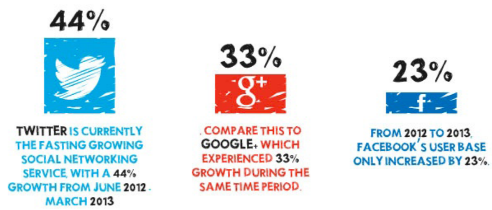
However, Twitter is qualified to be at the first place for the growth of active users in the last year, with an increase of 44%, followed by Google+ with an increase of 33% and Facebook with 23%. It should be noted that the growth of Google+ seems partly inflated by the need to register to it in order to use other services offered by Google, which owns a large slice of the market for web services. For example, in order to add comments on the video-sharing platform YouTube or to review applications on Android Market Google Plus you must be registered to Google+.

Regarding the average time spent by users on these three social media, the pride of place certainly belong to Facebook with 6 hours and 44 minutes (average time down with respect to 7 hours and 9 minutes of the previous year), followed by Twitter with 33 minutes. Google+ has less than 7 minutes. Finally, if we consider the geographical regions, we can see that in the first quarter of 2013 Facebook users have fallen in some countries including the United States (6 million active users less than the first quarter of 2013), and the UK (-1.4 million active users), while they increased especially in South America and Asia. The United States, however, remain the largest market of Facebook in the world with 113.4 million active users, followed by India

5 It is a measure of brand or category popularity. It is defined as the number of people who buy a specific brand or a category of goods at least once in a given period, divided by the size of the relevant market population (Farris et al., 2010)



(a) Visitor growth



(b) Growth rate

Figure 1: Social media stats (2004-2013). *Source:* searchenginejournal.com

(62.7 million), Brazil (58.5 million), Indonesia (51 million) and Mexico (38.4 million).

The countries that have registered in the last year the most absolute growth in the number of Twitter users were Indonesia, Saudi Arabia and Singapore, with rates between 44% and 35%. In terms of active users, the country with the greatest number of Twitter users is China, with 35.5 million, followed by India with 33 million and the U.S. with 22.9 million. The growth of Twitter active users is very high: +96% in Hong Kong, +93% in USA, +63% in Russia and +58% in China. Regarding Google+, the top three countries as the number of active users are China (100 million), India (over 40 million), Brazil and Indonesia (more than 20 million). Relatively low, however, its penetration rate in a market as the U.S. one (6%).[6]

2.2 ADVANTAGES OF ANALYSIS ON TWITTER

The reason for the popularity of Twitter is attributable to a number of factors. First, the percentage of public profiles, whose content is then

6 Stats from eMarketer.com, mashable.com, internetworldstats.com, socialbakers.com and globalwebindex.net

directly analyzed by the researchers through the streaming API (Application Programming Interface) provided by Twitter itself, is much higher than other social media. For example, in 2012, just over 11% of Twitter users were using the private profiles, compared to over 53% of Facebook. This turns Twitter into a gold mine of free data. On the other hand, by using Twitter you can also make a geolocation, i. e. the identification of the latitude and longitude of the Twitter's user when this decides to "post" a message (if the user has chosen to allow the access to the information related to its location). Third, the presence of hashtags gives more visibility to the messages that can be easily grasped from researchers and allows messages to be read by a wider audience, linking posts from users who do not know each other. It is not a coincidence that Twitter, despite the apparent constraint of 140 characters, not only has "the ability to drive the traffic of information among all the online platforms" more than other social channels (Parmelee and Bichard, 2011), but also the ability to identify what are the hot topics.

Twitter has several public profiles and hashtags gives more visibility to the messages which can also be geolocated

3 SENTIMENT ANALYSIS

"Everything we hear is an opinion,
not a fact.
Everything we see is a perspective,
not the truth."
— Marcus Aurelius

Opinions are central to almost all human activities and are key influencers of our behaviors. For this reason, when we need to make a decision we often seek out the opinions of others related to any kind of topic, such as which restaurant is the best one in the city, or who would they vote for in the next elections. This is important for individuals as well as for organizations. Today, thanks to Sentiment Analysis, this is no longer strictly necessary, especially when large amounts of data needs to be analyzed.
According to the definition reported in (Pang and Lee, 2008), sentiment suggests "*a settled opinion reflective of one's feelings*". The aim of Sentiment Analysis (SA), is therefore to define automatic tools able to extract subjective information from texts in natural language, such as opinions and sentiments, in order to create structured and actionable knowledge to be used by either a decision support system or a decision maker (Pozzi et al., 2013a). Sentiment Analysis is a term used to refer to the more general field of study that "*analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes*" (Liu, 2012).

Sentiment suggests a settled opinion reflective of one's feelings

Sentiment Analysis is often improperly used when referring to **Polarity classification** (Section 3.3), which instead is a sub-task aimed at extracting positive, negative or neutral sentiments (also called *polarities*) from texts. Although an opinion could also have a neutral polarity (e. g., "*I don't know if I liked the movie or not. I should watch it quietly.*"), most of the works in SA usually assume only positive and negative sentiments for sake of simplicity. Depending on the field of application, there are several names to refer to SA, e. g., Opinion Mining, Opinion Extraction, Sentiment Mining, Subjectivity Analysis, Affect Analysis, Emotion Analysis, Review Mining, etc. A taxonomy of the most popular SA tasks is reported in Figure 2.

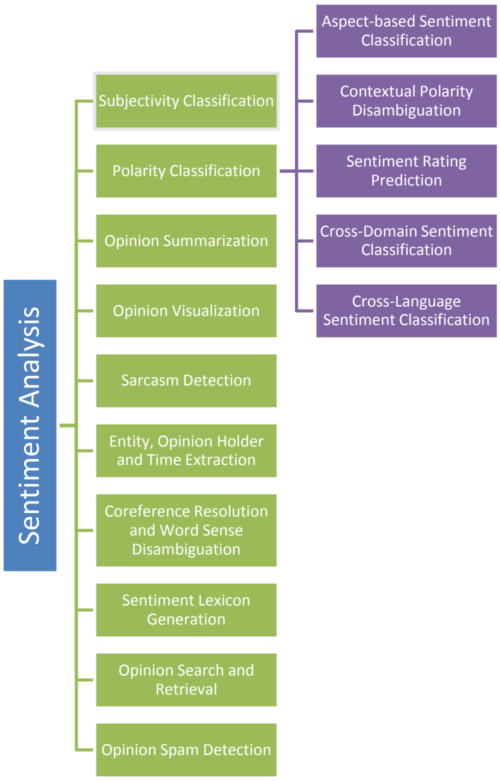


Figure 2: Sentiment Analysis tasks

3.1 TRADITIONAL SENTIMENT ANALYSIS

Sentiment Analysis is one of the most active research areas in NLP only since early 2000, when it has become a very active research area for different reasons (Liu, 2012). First, it has a wide array of applications, in a multitude of domains. Second, the challenging research problems it offers had never been studied in the past. Third, a huge volume of opinionated data recorded and easily accessible in digital forms is now available on the Web. In particular, the inception and rapid growth of the field coincide with those of the social media on the Web, e. g., reviews, forum discussions, blogs, microblogs (and consequently with social networks such as Twitter and Facebook). In fact, Sentiment Analysis is now the greatest focus of the social media research (as this thesis demonstrates. See Section 3.2).

SA is a multi-faceted and multidisciplinary field. Regarding computer science, it is widely studied in Data Mining, Web mining, and Text Mining. Due to its importance to business and society, it has spread from computer science to management and social sciences. In recent years, industrial activities surrounding SA have also thrived: numerous startups have emerged and many large corporations have built their own in-house capabilities (Liu, 2012).

3.1.1 *Sentiment Analysis formalization*

More formally, as defined in (Liu, 2012), an *opinion* is a quintuple,

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l), \tag{1}$$

where

- e_i is the name of an entity,
- a_{ij} is an aspect of e_i,
- s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i,
- h_k , and
- t_l is the time when the opinion is expressed by h_k.

The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength/intensity levels, such as the 1 to 5 stars system used by most review web-sites (e. g., www.amazon.com).
For example, consider that yesterday **John** bought an iPhone. He tested it during the whole day and when he go home from work (at **19:00** of **2-15-2014**) he wrote on his favorite social network the message "***iPhone** is very good, but they still need to work on **battery life** and **security** issues.*". Let us index 'iPhone', 'battery life' and 'security' as 1, 2 and 3, respectively. John is indexed as 4 and the time he wrote the sentence as 5. Then, John is the opinion holder h_4 and t_5 ('19:00 2-15-2014') is the time when the opinion is expressed by h_4 (John) . Term 'iPhone' is the entity e_1, 'battery life' and 'security issues' are aspects a_{12} and a_{13} of entity e_1 ('iPhone'), $s_{1245} = neg$ is the sentiment on aspect a_{12} ('battery life') of entity e_1 ('iPhone') and $s_{1345} = neg$ is the sentiment on aspect a_{13} ('security issues') of entity e_1 ('iPhone').
Please consider that when an opinion is on the entity itself as a whole, the special aspect 'GENERAL' is used to denote it.

Recall the definition of SA reported above: "*The aim of SA is therefore to define automatic tools able to extract subjective information in order to create **structured** and actionable knowledge [...]*". In line with this, the quintuple-based definition provides a framework to transform unstructured text to structured data (e. g., a database table). Then a rich set of qualitative, quantitative, and trend analyses can be performed using traditional database management systems (DBMS) and OLAP tools.

The quintuple-based representation provides a framework to transform unstructured text to structured data

3.1.2 *Applications*

One of the most important need of businesses and organizations in the real world is to find and analyze consumer or public opinions about their products and services (e. g., Why aren't consumers buying our laptop?). Knowing the opinions of existing users regarding a specific product is also interesting for individual consumers. This information could be useful to decide whether buying the product or not. This shows that decision making processes are also common in the everyday lives. However, with the advent of SA, an individual is no longer strictly limited to ask friends and family opinions or an organization limited to conduct surveys, opinion polls, and focus groups to sound out public or consumer opinions.

Sentiment Analysis paves the way to several and interesting applications, in almost every possible domain. For example, summarizing user reviews is a relevant task. In addition, errors in user ratings could be fixed (Pang and Lee, 2008): it is possible that users accidentally select a low rating when their review indicates a positive evaluation. Moreover, opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking (Goldberg et al., 2007; Hopkins and King, 2007). For instance, the USA president Barack Obama used SA to gauge feelings of core voters during the 2008 presidential elections. Other projects have as a long term goal the clarification of politicians' positions, such as what public figures support or oppose, to enhance the quality of information that voters have access to (Bansal et al., 2008; Greene, 2007). A further task is the augmentation of recommendation systems, where the system might not recommend items that receive several negative feedback (Pang and Lee, 2008). Moreover, ads are displayed sidebars in some online systems. It could be useful to detect web-pages that contain content inappropriate for ads placement (Jin et al., 2007). It could be useful to highlight product ads when relevant positive sentiments are detected, and hide the ads when negative statements are discovered. However, opinionated documents could also have the form of organizations' internal data, e. g., customer feedback. Sentiment Analysis applications have spread to several domains, from services and health care to financial services and political elections. There have been at least 40-60 start-up companies only in the USA. Big corporations such as Microsoft, Google, Hewlett-Packard, SAP, and SAS have also built their own in-house capabilities (Liu, 2012). Sentiment Analysis is an appealing field which leads companies/organizations and researchers to stimulate each other: practical applications and industrial interests have provided strong motivations for research in SA, and, on the other hand, the interesting and promising results in the SA research have given rise to strong interests and motivations in companies and organization.

However, Sentiment Analysis can also be applied to more ethical principles. For example, based on observations of Twitter's role in civilian response during the recent 2009 Jakarta and Mumbai terrorist attacks, (Cheong and Lee, 2011) proposed a structured framework to harvest civilian sentiment and response on Twitter during terrorism scenarios. Coupled with intelligent data mining, visualization, and filtering methods, this data can be collated into a knowledge base that would be of great utility to decision-makers and the authorities for rapid response and monitoring during such scenarios. Sentiment Analysis is also applied to the medical field. Cobb et al. (Cobb et al., 2013) applied Sentiment Analysis to examine how exposure to messages about the cessation drug varenicline (used to treat nicotine addiction) affects smokers' decision making around its use.

3.1.3 *Sentiment Analysis characteristics*

Sentiment Analysis is a broad and complex field of research, which carries with it the well-known issues of NLP and gives rise to new and interesting challenges. In the following, the main characteristics that constitute SA are described and discussed in details.

3.1.3.1 *NLP issues*

Sentiment Analysis touches almost every aspect of NLP, e. g., coreference resolution, negation handling, word sense disambiguation, Named-Entity-Recognition and Part-Of-Speech extraction which add more difficulties since these are still open issues in NLP. However, SA is not strictly a NLP problem because the fully understanding of the semantics of each piece of text is not needed. It "only" needs to understand some aspects of a piece of text, i. e. positive or negative sentiments and their target entities. In this sense, SA offers a great opportunity to make effective progresses both for SA itself and NLP (Liu, 2012).

Sentiment Analysis is not strictly a NLP problem

3.1.3.2 *Sentiment categorization: Objective vs. Subjective sentences*

The first aim when dealing with SA usually consists in distinguish between subjective and objective sentences. If a given sentence is classified as objective, no other fundamental tasks are required, while if the sentence is classified as subjective, its polarity (positive, negative or neutral) needs to be estimated (Figure 3). Subjectivity classification (Wiebe et al., 1999) is the task which distinguishes sentences that express objective (or factual) information (i. e. **objective sentences**) from sentences that express subjective views and opinions (i. e. **subjective sentences**). An example of objective sentence is "*iPhone is a smartphone*", while an example of subjective sentence is "*iPhone is awesome*". Polarity classification is the task which distinguishes sentences that express positive, negative or neutral polarities. Please note that a subjective sentence may not express any positive or negative sentiment (e. g., "*I guess he is arrived*"). For this reason, it should be classified as 'neutral'.

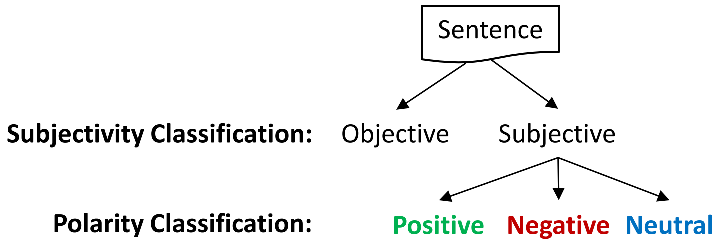


Figure 3: Sentiment Analysis workflow

3.1.3.3 *Levels of Analysis*

As reported in Section 3.1, the aim of SA is to "*define automatic tools able to extract subjective information* ***from texts*** *in natural language [...]*". The first choice when applying SA is to define what 'text' (i. e. the analyzed object) means in the considered case of study. In general, SA can be investigated mainly at three levels (graphically represented in Figure 4):

- **Document-level**: the aim at this level is to classify the polarity of a whole opinionated document. For example, given a product review, the system determines whether the review expresses an overall positive, negative or neutral opinion about the product. This task is commonly known as document-level sentiment classification. The assumption is that each document expresses only one opinion on a single entity (e. g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

- **Sentence-level**: the aim at this level determines the polarity of each sentence of a document. This task is commonly known as sentence-level sentiment classification. The assumption is that each sentence expresses only one opinion on a single entity. Thus, similarly to document-level sentiment classification, it is not applicable to sentences which evaluate or compare multiple entities.

- **Entity and Aspect-level**: performs a finer-grained analysis than document and sentence level. It is also the finest possible analysis applicable to texts. It is based on the idea that an opinion consists of a *sentiment* and a *target* (of opinion). In other words, it discovers pairs of *{aspect, sentiment}*. An opinion without its target is of limited use. For example, the sentence "*iPhone is very good, but they still need to work on battery life and security issues.*" evaluates three aspects: iPhone, battery life and security. The sentiment on iPhone is positive, but the sentiment on its battery life and security is negative. Although both document and sentence level sentiment classifications are already highly challenging, the analysis performed at aspect-level is even more difficult (Liu, 2012).

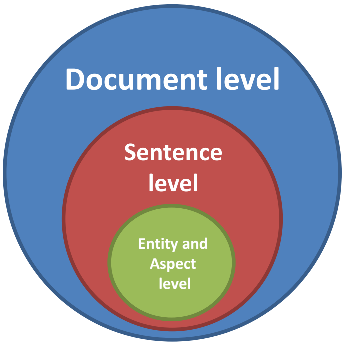


Figure 4: Different Levels of Analysis

3.1.3.4 *Regular vs. comparative opinion*

An opinion can assume different shades and can be assigned to one of the following groups[1]:

- **Regular opinion**: a regular opinion is often referred in the literature to a standard opinion and it has two main sub-types:
  - **Direct opinion**: a direct opinion refers to an opinion expressed directly on an entity (e. g., "*The **screen brightness** of the iPhone is awesome.*")
  - **Indirect opinion**: an indirect opinion is an opinion that is expressed indirectly on an entity based on its effects on some other entities. For example, the sentence "*After I switched to iPhone, I lost all my data!*" describes an undesirable effect of the switch on 'my data', which indirectly gives a negative sentiment to iPhone.
- **Comparative opinion**: a comparative opinion expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities (Jindal and Liu, 2006). For example, the sentences, "*iOS is more performant than Android*" and "*iOS is the most performing operating system*" express two comparative opinions. A comparative opinion is usually expressed using the comparative or superlative form of an adjective or adverb.

3.1.3.5 *Explicit vs. Implicit Opinions*

Among the different shades that an opinion can assume, we have to distinguish explicit and implicit opinions:

1 Definitions from (Liu, 2012)

- **Explicit opinion**: an explicit opinion is a subjective statement that gives a regular or comparative opinion (e. g., "*The screen brightness of the iPhone is awesome.*").

- **Implicit opinion**: an implicit opinion is an objective statement that implies a regular or comparative opinion that usually expresses a desirable or undesirable fact, e. g., "*Saturday night I'll go to the cinema to watch 'Lone survivor'. I cannot wait to watch it!*" and "*'Saving Soldier Ryan' is more violent than 'Lone survivor'*". The first example suggests that there is some good expectation about he movie, although it is not explicated in words, while the understanding of the hidden opinion in the second example is difficult even for humans. For some people, violence in war movies could be a good characteristic that makes the movie more realistic, while a negative feature for others.

Clearly, explicit opinions are easier to detect and to classify than implicit opinions. Much of the current research has focused on explicit opinions. Relatively less work has been done on implicit opinions (Zhang and Liu, 2011).

3.1.3.6 *Dealing with Figures of Speech*

A figure of speech is any artful deviation from the ordinary mode of speaking or writing

Corbett defines a figure of speech as any artful deviation from the ordinary mode of speaking or writing. In the tradition of Aristotle, (Corbett, 1971) divides figures of speech into two groups: *schemes* and *tropes*. According to Corbett, the function of schemes and tropes is to carry out a transference of some kind; schemes are characterized by a transference in order, while tropes are characterized by a transference in meaning.

For example, the most problematic figures of speech in NLP are **irony** and **sarcasm**, which are collocated under the tropes group. While irony is often used to emphasize occurrences that deviate from the expected, such as twists of fate, sarcasm is commonly used to convey implicit criticism with a particular victim as its target (McDonald, 1999). Examples of sarcastic and ironic sentences are:

1. **Sarcasm**
   (*Note: Alice hates Bill's travel books*)
   – Alice: Yeah, I like, really dig your travel books, Bill. You're a really skillful author.
   – Bill: Oh.

2. **Irony**
   (*Note: Bill and Alice have just seen a really appalling play at the theater. Both Bill and Alice are disappointed.*)
   – Bill: Well! What a worthwhile use of an evening!
   – Alice : Yeah.

In the irony example, there was no sarcasm because Bill was not intending to wound Alice with his comment. He was using irony to remark that he felt he had wasted his evening at the theater. In the sarcasm example, indeed, Alice used sarcasm to show Bill that she did not like his books and thought that he is not a good writer. There's irony too, but the tone of the delivery which convey implicit criticism makes it sarcastic.

While irony is often used to emphasize occurrences that deviate from the expected, sarcasm is commonly used to convey implicit criticism

One inherent characteristic of the sarcastic and irony speech acts is that they are sometimes hard to recognize, fist for humans and then for machines. The difficulty in the recognition of sarcasm and irony causes misunderstanding in everyday communication and poses problems to many NLP systems due to the poor results obtained by state-of-the-art works. In the context of SA (where sarcasm and irony are usually considered as synonyms) when a sarcastic/ironic sentence is detected as positive, it means negative, and vice-versa.

In (Davidov et al., 2010), a semi-supervised learning approach was proposed to identify sarcasms. It automatically expanded the seed set (a small set of labeled sentences) through Web search. The authors posited that sarcastic sentences frequently co-occur in texts with other sarcastic sentences. An automated web-search was performed using each sentence in the seed training set as a query. The system then collected up to 50 search engine snippets for each seed example and added the collected sentences to the training set. This enriched training set was then used for learning and classification. For learning, it used two types of features: pattern-based and punctuation-based features. A pattern is simply an ordered sequence of high frequency words. Punctuation-based features include the number of '!', '?' and quotes, and the number of capitalized/all capital words in the sentence. This work, however, did not perform sentiment classification: it only separated sarcastic and non-sarcastic sentences.

In particular, (González-Ibáñez et al., 2011) studied the problem in the context of SA using Twitter data, i.e. they distinguish sarcastic tweets and non-sarcastic tweets that directly convey positive or negative opinions. Again, a supervised learning approach was taken using Support Vector Machines (SVM) and logistic regression. As features they used unigrams and some dictionary-based information. The dictionary-based features include (1) word categories; (2) WordNet Affect (WNA); and (3) a list of interjections (e.g., ah, oh, yeah), and punctuations (e.g., !, ?). Features like emoticons, and ToUser (which marks if a tweet is a reply to another tweet, signaled by <@user>) were also used. Experimental results for three-way classification (sarcastic, positive and negative) showed that the problem is very challenging: the best accuracy was only 57%.

3.2 SENTIMENT ANALYSIS ON SOCIAL MEDIA

Established that the current technological progresses is able nowadays to enable an efficient storing and retrieval of huge amount of data, the attention is nowadays on methodologies for extracting information and creating knowledge from raw sources. Social Media represent an emerging challenging sector in the context of Big Data: the natural language expressions of people can be easily reported through blogs and short text messages, rapidly creating unique contents of huge dimensions that must be efficiently and effectively analyzed to create actionable knowledge for decision making processes. The massive quantity of continuously contributing texts, which should be processed in real time in order to take informed decisions, calls for two main radical progresses: (1) a change of direction in the research through the transition from data-constrained to data-enabled paradigm (Gopal et al., 2011) and (2) the convergence to a multidisciplinary area that mainly takes advantage of psychology, sociology, natural language processing and machine learning.

A potential leverage towards novel decision support systems is represented by the transformation of qualitative data from user-generated contents to quantitative information when making decisions. In this context, the extraction of this subjective information is crucial to create structured and actionable knowledge to be used by either a decision support system or a decision maker. The knowledge embedded in user-generated contents has been shown to be of paramount importance from both user and company/organization points of view: people express opinions on any kind of topic in an unconstrained and unbiased environment, while corporations and institutions can gauge valuable information from raw sources. In order to make qualitative textual data effectively functional for decision processes, the quantification of "what people think" becomes a mandatory step. This issue is usually approached as a polarity detection task aimed at classifying texts as positive, negative or neutral.

Although there is a strong literature regarding the analysis of well-formed documents (e. g., newspaper articles, reviews, official news, etc.) as presented in Section 3.1, there are still many challenges to be faced on social media in order to get a feel for what people think about current topics of interest. In order to make best use of these recent and highly dynamic source of information, we need to be able to distinguish what is important and interesting. There are obvious benefits to companies, governments and, more in general, organizations in understanding what the public think about their products and services. The spread of information through social networks can also trigger a chain of reactions to such situations and events which ultimately lead to administrative, political and societal changes. Social media users generate content that is dynamic, rapidly changing to reflect the societal and sentimental fluctuations of the authors as well as the ever-changing use of language.

Considering the evolution of the sources where opinions are issued, the strategies available in the current state of the art are therefore no longer effective for mining opinions in this new and challenging environment. In fact, Social Media Sentiment Analysis (SM-SA), in addition to inherits a multitude of issues coming from traditional SA and NLP (Figure 5), introduces new and complex challenges.

The strategies available in the current state of the art are therefore no longer effective for mining opinions on Social Media

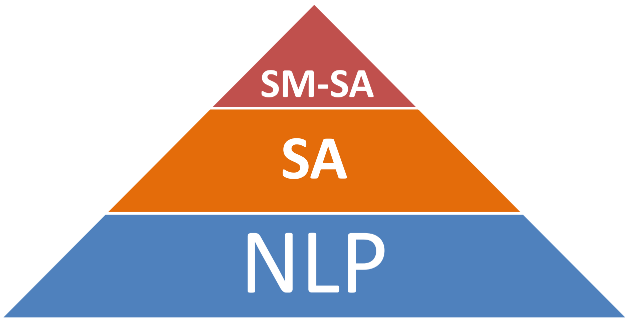


Figure 5: Dependency levels among NLP, SA and SM-SA

Microposts such as tweets are, in some sense, the most challenging text type for text mining techniques, and in particular for SA, due to different reasons:

- **Short messages (microtexts)**: Twitter and most Facebook messages are very short (max 140 characters for tweets). Many semantic-based methods supplement this lack with extra information and context coming from embedded URLs and hashtags. For instance, (Abel et al., 2011) augment tweets by linking them to contemporaneous news articles, whereas (Mendes et al., 2010) exploit online hashtag glossaries to augment tweets.

- **Noisy content**: unlike well-formed documents (e. g., news or reviews) which usually have a well-defined linguistic form, Social Media tend to be less grammatical, due to an excessive amounts of colloquialism, with hasty spelling (e. g., 2u) and unconventional grammatical quirks. Moreover, social media posts are full of irregular capitalisation (e. g., all capital or all lowercase letters), emoticons (e. g., :-P), idiosyncratic abbreviations (e. g., ROFL, ZOMG) and "word lengthening", i. e. replication of letters as if words were screamed in order to emphasize the sentiment (e. g., "noooooooo", "I'm haaaappyyyyyyy", etc.).
Spelling and capitalisation normalisation methods have been developed (Han and Baldwin, 2011), coupled with studies of location-based linguistic variations in shortening styles in microtexts (Gouws et al., 2011). Emoticons are used as strong sentiment indicators in opinion mining algorithms.

- **Time**: in addition to linguistic analysis, social media content lends itself to analysis along temporal lines, which is a relatively under-researched problem. Addressing the temporal dimension of social media is a pre-requisite for much-needed models of conflicting and consensual information, as well as for modeling change in user interests. Moreover, temporal modeling can be combined with opinion mining, to examine the volatility of attitudes towards topics over time.

- **Social context**: it is crucial for the correct interpretation of social media content. Semantic-based methods need to make use of social context (e. g., who is the user connected to, how frequently they interact, etc.), in order to derive automatically semantic models of social networks, measure user authority, cluster similar users into groups, as well as model trust and strength of connection.

- **Metadata**: since users produce, as well as consume social media content, there is a rich source of explicit and implicit information about the user, e. g., demographics (gender, location, age, etc.), interests, opinions, that can be used to improve classification performance.

- **Multilingual**: social media content is strongly multilingual. For instance, less than 50% of tweets are in English, with Japanese, Spanish, Portuguese, and German also featuring prominently (Carter et al., 2013). Semantic technology methods have so far mostly focused on English, while lowoverhead adaptation to new languages still remains an open issue. Automatic language identification (Carter et al., 2013; Baldwin and Lui, 2010) is an important first step, allowing applications to first separate social media in language clusters, which can then be processed using different algorithms.

If well handled, these characteristics can form an important part of the meaning. For this reason, although less organized, such information should also be considered as a gold mine which can not be handled in well-formed documents. Another challenge is that Social Media contain highly diverse and often latest topics.

In general, sentiment classification techniques for traditional SA can be roughly divided into **lexicon-based methods** (Scharl and Weichselbraun, 2008; Taboada et al., 2011) and **Machine Learning methods** (Boiy and Moens, 2009). Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine Learning approaches make use of syntactic and/or linguistic features, and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods (Hu and Liu, 2004). For example, (Moghaddam and Popowich, 2010) establish the polarity of reviews by identifying the polarity of the adjectives that appear in them, with a reported accuracy of about 10% higher than pure machine learning techniques.

Following the traditional line of standard SA, most of the works regarding SM-SA has predominantly used Machine Learning techniques. One of the reasons for the relative paucity of ad-hoc linguistic techniques for SM-SA is most likely due to the difficulties in using NLP on low quality text, something which machine learning techniques can, to some extent, bypass with sufficient training data. (Pak and Paroubek, 2010) aimed at classifying arbitrary tweets on the basis of positive, negative and neutral sentiment, constructing a classifier which used n-gram and POS features. Their approach has much in common with an earlier sentiment classifier constructed by (Go et al., 2009). The work of (Ritter et al., 2011) demonstrated some of the difficulties in applying traditional POS tagging, chunking and Named Entity Recognition techniques to tweets, proposing a solution based on Labeled-LDA (Ramage et al., 2009).

Most of the works regarding Social Media Sentiment Analysis has predominantly used Machine Learning techniques

However, all these approaches are consistent with the classical statistical inference problem formulation, in which instances are homogeneous, independent and identically distributed (i.i.d. assumption). In other words, they merely considered textual information only, not taking into account that Social Media, the today's largest and richest sources of information, are actually networked environments. Although relationship information of social networks has been extensively investigated, the work of incorporating content and relationship information to facilitate polarity detection has not been thoroughly studied yet (some exceptions are (Tan et al., 2011; Hu et al., 2013; Rabelo et al., 2012)). This problem is at the heart of the recent literature on SA applied to Online Social Media.

The work of incorporating content and relationship information to facilitate polarity detection has not been thoroughly studied yet

3.3 Polarity Classification

Sentiment Analysis is composed of a sub-task called "*Polarity Classification*", also known as "Polarity Detection" or "Sentiment Classification". It is aimed at extracting positive, negative or neutral sentiments (also called polarities) from texts. Although an opinion could also have a neutral polarity (e. g., "*I don't know if I liked the movie or not. I should watch it quietly*"), most of the works in SA usually assume only positive and negative sentiments for sake of simplicity. The polarity classification task can be addressed with supervised, unsupervised and semi-supervised techniques.

3.3.1 *Learning by supervision*

Sentiment classification can be treated as a text classification problem where, instead of classifying documents of different topics (e. g., politics, sciences, and sports), "positive", "negative" and "neutral" classes are estimated. Online reviews and social media posts are usually used as training and testing data. However, supervised learning problems are by definition characterized by a tremendous human effort in labeling data which typically grows with the amount of data available. Since online reviews have rating scores assigned by their reviewers (e. g., 1-5 stars), this human effort can be avoided discretizing the ratings (e. g., 4 or 5 stars is positive, 1 and 2 stars is negative). Most research papers do not use the neutral class, which would makes the classification problem considerably harder. In traditional classification problems, topic related words are the key features. However, in polarity classification, sentiment words that are more prone to indicate positive or negative opinions are used (e. g., great, excellent, wow, etc.). Since it is a text classification problem, any existing supervised learning method, such as Naïve Bayes classifier and Support Vector Machines (SVM) (Joachims, 1999; Cristianini and Shawe-Taylor, 2000), can be easily applied. It was shown in (Pang et al., 2002) that using unigrams as features for sentiment classification of movie reviews performed quite well with either Naïve Bayes or SVM, although the authors also tried a number of other feature options.

Although other paradigms have been investigated, supervised learning still remains an active approach for polarity classification, thanks to its ability in predicting sentiment orientations. But in subsequent research, many more features and learning algorithms were tried by a large number of researchers. Some of these features are:

- **Part-of-Speech**: words of different Parts-Of-Speech (POS) may be treated differently. For example, it was shown that adjectives are important indicators of opinions (Mullen and Collier, 2004; Whitelaw et al., 2005). However, all the POS tags and their n-grams can be used as features.

- **Sentiment words**: words that are used to express positive or negative sentiments, which are usually contained in sentiment lexicons (see for example (Hu and Liu, 2004)).

- **Sentiment shifters**: these are expressions that are used to reverse the sentiment orientations, i. e. from positive to negative or vice-versa. Negation words are the most important class of sentiment shifters. For example, the sentence "I don't like this camera" is negative, although 'like' is positive. Please note that not all occurrences of such words mean sentiment inversions (e. g., "*not only... but also...*").

- **Expressive signals**: typically used in online social media, expressive signals can be used to expand the feature set. The most common expressive signals are: *emoticons* (':-)', '=)', 'XD', ':('), *emphatic* ('ROFL', 'BM') and *onomatopoeic* ('bleh', 'wow') *expressions* and *expressive lengthening*, i. e. indication of emphasis that is strongly associated with subjectivity and sentiment, such as *"I'm happyyyyyyy!!"* or *"Nooooooooo"*. (Pozzi et al., 2013b; Fersini et al., 2014a) investigated the impact of expressive signals in respect of polarity classification of tweets by proposing a combination of language normalization and feature expansion. The considered expressive signals has been used to enrich the feature space and train several baseline and ensemble classifiers aimed at polarity classification.

SOME SUPERVISED CLASSIFIERS In the following, the most popular supervised classifiers for polarity classification are briefly presented:

- *Naïve Bayes (NB)* (McCallum and Nigam, 1998) is the simplest generative model that can be applied to the polarity classification task. It predicts the polarity label $l^*(s)$ of a sentence s given a vector representation of textual cues $t \in V$ by exploiting the Bayes' Theorem:

$$P(l(s) = k \mid t_1, t_2, \dots, t_{|V|}) = \frac{P(k)P(t_1, t_2, \dots, t_{|V|} \mid l(s) = k)}{P(t_1, t_2, \dots, t_{|V|})} \quad (2)$$

  Thus, the polarity label $l^*(s)$ is determined according to the following maximum a posteriori (MAP) decision rule:

$$l^*(s) = \arg\max_k P(l(s) = k) \prod_{v=1}^{|V|} P(t_v \mid l(s) = k) \quad (3)$$

  For training NB, a multinomial distribution has been assumed.

- *Support Vector Machines (SVM)* (Cortes and Vapnik, 1995) are learning machines that try to find the optimal hyperplane discriminating samples of different classes. In order to take advantage of SVM predictions when the model is enclosed in an ensemble, probabilistic SVM have been adopted. To this purpose, the following optimization problem has been solved by using an improved implementation (Lin et al., 2007) of (Platt, 2000):

$$\min_p \quad \frac{1}{2} \sum_{k=1}^{|K|} \sum_{k'=1:k' \neq k}^{|K|} (\tau_{k'k} p_k - \tau_{kk'} p_{k'})^2 \quad (4)$$

$$\text{subject to} \quad p_k \geqslant 0, \forall k, \quad \sum_{k=1}^{|K|} p_k = 1 \quad (5)$$

  where the decision variable p_k denotes the probability of a given message to be labeled as k and $\tau_{kk'}$ is the pairwise class probabilities. The optimal label $l^*(s)$ is determined by selecting the label with the highest probability p_k. For training SVM, a linear kernel has been assumed.

- *Maximum Entropy (ME)* is a classifier largely adopted for the sentiment analysis tasks (McCallum et al., 2006). The over-riding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible and therefore have a maximal entropy. This classifier is based on the assumption that the log probability of each label $l(s)$ is a linear combination of feature functions F. Hence, the conditional distribution of a label $l(s)$ is defined as follows:

$$p(l(s) = k) = \frac{1}{Z(F)} \sum_{o=1}^{|F|} \exp\{\lambda_o f_o(s)\} \tag{6}$$

where λ_o is the weight associated to the feature f_o, $Z(F)$ is a normalization factor that ensures a consistent probability distribution of $l(s)$. In this study, the ME model is trained with feature functions that represent the unigrams within a sentence. The polarity label $l^*(s)$ is selected according to the maximum likelihood $P(l(s) = k)$:

$$l^*(s) = \arg\max_k P(l(s) = k) \tag{7}$$

- *Conditional Random Fields (CRF)* (Sutton and McCallum, 2012) are discriminative probabilistic graphical models able to directly learn the conditional distribution $p(k|s)$ by taking into account the sequential patterns of the sentences. Formally, a Conditional Random Fields is defined as:

$$p(l(s) = k) = \frac{1}{Z(s)} \exp\left\{\sum_u \lambda_u g_u(s) + \sum_o \mu_o h_o(s)\right\} \tag{8}$$

where g_u and h_o are state and transition feature functions, λ_u and μ_o are the corresponding weights and $Z(s)$ is the observation-dependent normalization factor.

According to Equation (8), CRF results to be globally conditioned on the observation s. The simplest form of graph we can think of is a linear chain, which is nevertheless very useful to model a sequence of observations making up a sentence. Sentences do possess a somewhat linear nature, as they are sequences of words. The optimal polarity label $l^*(s)$ is selected according to the maximum likelihood defined in equation Equation (8).

3.3.2 *Learning without supervision*

As mentioned in the previous section, sentiment words are often the key factor for sentiment classification. For this reason, sentiment classification performed in an unsupervised manner through lexicons is straightforward. Several unsupervised learning approaches follow two main steps: (1) create a sentiment lexicon in an unsupervised manner, and (2) determine the degree of positivity (or negativity) of a text unit via some function based on the positive and negative indicators, as determined by the lexicon (Hatzivassiloglou and Wiebe, 2000; Yu and Hatzivassiloglou, 2003). Some interesting variants of this general technique consist in using the polarity of the previous sentence as a tie-breaker when the scoring function does not indicate a definitive classification of a given sentence (Hu and Liu, 2004), or to incorporate information drawn from some labeled data (Beineke et al., 2004).

The method proposed in (Turney, 2002) is one of the most famous unsupervised technique for polarity classification. It performs classification based on some fixed syntactic patterns, based on Part-Of-Speech (POS), that are likely to be used to express opinions.

The algorithm consists of three steps:

1. The main assumption underlying the proposed method is that adjectives (JJ), adverbs (RB), comparative adverbs (RBR) and superlative adverbs (RBS) often express opinions. For this reason, two consecutive words are extracted if their POS tags conform to any of the patterns in Table 1. For example, the first pattern means that two consecutive words are extracted if the first word is an adjective, the second word is a singular (NN) or plural (NNS) noun, and the third word (not extracted) can be anything. For example, in the sentence "The iPhone has a bright screen", "bright screen" is extracted as it satisfies the first pattern.

2. It estimates the Sentiment Orientation (SO) of the extracted phrases through Pointwise Mutual Information (PMI) measure, which measures the degree of statistical dependence between two terms:

$$\mathrm{PMI}(\mathrm{term}_1, \mathrm{term}_2) = \log_2\left(\frac{P(\mathrm{term}_1 \wedge \mathrm{term}_2)}{P(\mathrm{term}_1)P(\mathrm{term}_2)}\right) \tag{9}$$

   where $P(\mathrm{term}_1 \wedge \mathrm{term}_2)$ is the actual co-occurrence probability of term_1 and term_2, and $P(\mathrm{term}_1)P(\mathrm{term}_2)$ is the co-occurrence

	First word	**Second word**	**Third word** (not extracted)
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table 1: Patterns of POS tags for extracting two-word phrases

probability of the two terms if they are statistically independent. The log of this ratio is the amount of information that we acquire about the presence of one of the words when we observe the other. The Sentiment Orientation (SO) of a phrase is computed as follows:

$$\begin{aligned} SO(phrase) = PMI(phrase, \text{"excellent"}) \\ - PMI(phrase, \text{"poor"}) \end{aligned} \tag{10}$$

The reference words "excellent" and "poor" were chosen because, in the five star review rating system, it is common to define one star as "poor" and five stars as "excellent". SO is positive when phrase is more strongly associated with "excellent" and negative when phrase is more strongly associated with "poor".

By searching the two terms together and separately, the probabilities in Equation (9) are calculated by issuing queries to the web search engine AltaVista and collecting the number of hits. AltaVista was used because it has a NEAR operator to constrain the search to documents that contain the words within a window of ten words in either order. For this reason, Equation (11) can be rewritten as:

$$\begin{aligned} &SO(\text{phrase}) = \\ &\log_2\left(\frac{hits(\text{phraseNEAR"excellent"})hits(\text{"poor"})}{hits(\text{phraseNEAR"poor"})hits(\text{"excellent"})}\right) \end{aligned} \tag{11}$$

where $hits(query)$ is the number of hits returned.

3. Given a document (e. g., review), the average SO of all phrases in the document is computed. The document is classified as positive if the average SO is positive and negative otherwise.

Another very popular unsupervised approach is the *Dictionary-based classifier (DIC)*, the simplest and naive method for the polarity classification task. Given two dictionaries of positive and negative terms (e. g., DictHuLiu[2] (Hu and Liu, 2004) or SentiWordNet[3] (Esuli and Sebastiani, 2006)) the sentence polarity is determined, checking if each sentence term belongs to the positive or the negative dictionary. The final label is inferred through a majority voting approach (terms that are not included neither in the positive nor in the negative dictionary are not considered during the aggregation process). The output of DIC can be easily modified to be probabilistic: given the set K of labels to be predicted, the a posteriori probability for label $l(s) = k$ is estimated as the number of terms labeled as $k \in K$ over the total number of positive or negative terms within the sentence.

2 www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
3 sentiwordnet.isti.cnr.it

3.3.3 *Learning by semi-supervision*

Although supervised learning methods have been widely employed and proven effective in sentiment classification in the literature (Pang and Lee, 2008), they normally depend on a large amount of labeled data, which usually involves high cost in labor and time. On the other hand, acquisition of unlabeled data is relatively inexpensive but they usually produce lower performance. Falling between unsupervised and supervised learning, semi-supervised learning methods are proposed to effectively utilize a small scale of labeled data along with a larger amount of unlabeled data.

One kind of semi-supervised methods for polarity classification is to utilize prior lexical knowledge in conjunction with labeled and unlabeled data. For example, (Gamon et al., 2005) proposed a semi-supervised learning algorithm to learn from a small and a large set of labeled and unlabeled sentences respectively. The learning algorithm was based on Expectation-Maximization (EM) using the Naïve Bayes as the base classifier (Nigam et al., 2000). This work performed three-class classification: positive, negative, and "other" (no opinion or mixed opinion). In (Rao and Ravichandran, 2009), three graph-based semi-supervised learning methods were tried to separate positive and negative words given a positive and a negative seed set, and a synonym graph extracted from WordNet[4], a large lexical database of English words. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

Another kind of semi-supervised methods for polarity classification consists in employing some bootstrap techniques, such as self-training (Yarowsky, 1995) and co-training (Blum and Mitchell, 1998). Among them, co-training has been proven more effective than self-training. For example, (Wan, 2009), focusing on the problem of cross-lingual sentiment classification, proposed a co-training approach to overcome the lack of chinese sentiment corpora. It used an available English corpus as training data for Chinese sentiment classification.

Semi-supervised learning is often leveraged when dealing with networked environments

Semi-supervised learning is often leveraged when dealing with networked environments. The general idea is to start with some knowledge (labeled data) and exploit the network to classify the unlabeled data by a propagation effect. For example, (Tan et al., 2011) and (Pozzi et al., 2013c) proposed two semi-supervised frameworks that, given a small labeled set of messages, estimate user polarities by combining post contents and connections between authors. Thanks to the network, the prediction of each user has impact on all the other nodes by a propagation effect, smoothing each predicted label according to adjoining nodes. (Deng et al., 2013) proposed a novel information network-based framework which can infer hidden similarity and dis-

4 wordnet.princeton.edu

similarity between pair of users by exploring similar and opposite opinions, so as to improve post-level and user-level sentiment classification simultaneously.

3.3.4 *Learning in relational environments*

A key aspect of several online social networks is that they are rich in content and relationships, and provide unprecedented challenges and opportunities from the perspective of Sentiment Analysis, and subsequently knowledge discovery, data mining, social network analysis and text mining.

There are mainly three kinds of analyses in the context of social networks:

- **Link-based Analysis**: structural relationships are used to simulate and analyze the diffusion model of opinions across the network. For instance, (Gatti et al., 2014) presents a method to model and simulate interactive behavior in microblogs taking into account the users sentiment. The work presented in (Kaschesky and Riedl, 2011) aims at identifying and tracking how users join and construct opinions, and how some of these opinions around a specific topic spread.

- **Content-based Analysis**: in addition to relationships, social networks contain an extraordinary amount of contents which can be leveraged. For example, Message Networks (e. g., Facebook and Twitter) contain tons of unstructured text messages that can be analyzed by Natural Language Processing and Text Mining techniques for Sentiment Analysis tasks, while Multimedia Networks (e. g., Flickr and Youtube) allows users to add and share multimedia data coupled with short texts potentially opinionated. In these contexts, several machine learning approaches have been proposed in the literature (Pang and Lee, 2008).

- **Combining Link-based and Content-based Analysis**: although relationship and content information have been independently investigated in the past, their combination has not been thoroughly studied yet (some exceptions are (Tan et al., 2011; Hu et al., 2013; Rabelo et al., 2012; Speriosu et al., 2011)). This is particularly useful when textual features do not always provide sufficient information to infer sentiment (e. g., *I agree!*).

However, most of the works regarding sentiment classification usually take into account text as unique information to infer sentiment (Wang and Manning, 2012; Maas et al., 2011; Go et al., 2009; Barbosa and Feng, 2010). For instance, (Go et al., 2009) presented the results of machine learning algorithms for classifying the sentiments of Twitter messages using distant supervision (i. e. emoticons are considered as ground truth for tweet labelling), while (Barbosa and Feng, 2010) explored the linguistic characteristics of how tweets are written and the meta-information of words for sentiment classification. However, they merely considered textual information, do not taking into account that Social Networks are actually networked environments and that this additional information should be considered as a gold mine. For example, (Speriosu et al., 2011) proposed to enrich the content representation by including user followers as additional features. In (Tan et al., 2011; Hu et al., 2013), authors estimate user-level sentiment by exploring tweet contents and friendship relationships among users. However, in (Tan et al., 2011) tweet sentiment is not estimated but always assigned by matching the real user label, and the mention network (a user mentions another one using the Twitter @-convention) is also considered following the motivation that users will mention those who they agree with. Taking into account relationships could be useful also when dealing with implicit (or implied) opinions, where textual features do not always provide explicit information about sentiment. An implicit opinion is formally an objective statement that implies a regular or comparative opinion that usually expresses a desirable or undesirable fact, e. g., "*Saturday night I'll go to the cinema to watch 'Lone survivor'. I cannot wait to watch it!*" and "*'Saving Soldier Ryan' is more violent than 'Lone survivor'*". The first example suggests that there is some good expectation about the movie, although it is not encoded in words, while the understanding of the hidden opinion in the second example is difficult even for humans. For some people, violence in war movies could be a good characteristic that makes the movie more realistic, while a negative feature for others. Much of the current research has focused on explicit opinions because are easier to detect and classify than implicit opinions. Less work has been done on implicit opinions (Zhang and Liu, 2011), based on the consideration that textual features do not always provide explicit information about sentiment (e. g., *I cannot wait to watch it!*). Also in such cases, additional information, such as user network, needs to be considered. It has been observed that combining content-based and link-based analysis provides more effective results in a wide variety of applications (Fersini et al., 2010b; Pozzi et al., 2013c; Mei et al., 2008a). The combination of contents and relationships is a core task towards novel system of Sentiment Analysis.

It has been observed that combining content-based and link-based analysis provides more effective results

3.3.5 *Performance Measures*

As mentioned in Section 3.3.1, polarity classification is essentially a text classification problem usually formulated as a two-class classification (positive and negative). For this reason, the performance measures used for polarity classification are the classical state-of-the-art measures.

Let $\mathcal{D}$ be the testing set comprising n points, $K = \{pos, neg\}$ be the set of polarity labels, and i be a classifier. For $j \in \mathcal{D}$, let $l_i(j)$ denote its true class, and let $l_i^*(j)$ denote its predicted class.

- **Error Rate**: The error rate of a classifier i is the fraction of incorrect predictions over the testing set, defined as

$$\mathrm{ErrorRate} = \frac{1}{n} \sum_{j=1}^{n} I(l_i(j) \neq l_i^*(j)) \tag{12}$$

  where $I(\cdot)$ is an indicator function that has the value 1 when its argument is true, and 0 otherwise. Error rate is an estimate of the probability of misclassification. The lower the error rate the better the classifier.

- **Accuracy**: the accuracy of a classifier i is the fraction of correct predictions over the testing set

$$\mathrm{Acc} = \frac{1}{n} \sum_{j=1}^{n} I(l_i(j) = l_i^*(j)) = 1 - \mathrm{ErrorRate} \tag{13}$$

  Accuracy gives an estimate of the probability of a correct prediction, thus, the higher the accuracy, the better the classifier.

- **Accuracy/Precision**: the class-specific accuracy or precision of the classifier i for class $k \in \{pos, neg\}$ is given as the fraction of correct predictions as k over all points predicted to be in class k

$$\mathrm{Acc}_k = P_k = \frac{\sum_{j=1}^{n} I(l_i^*(j) = k, l_i^*(j) = l_i(j))}{\sum_{j=1}^{n} I(l_i^*(j) = k)} \tag{14}$$

  The higher is the accuracy on class k the better is the classifier.

- **Coverage/Recall**: the class-specific coverage or recall of i for class k is the fraction of correct predictions over all points in class k

$$\mathrm{Coverage}_k = R_k = \frac{\sum_{j=1}^{n} I(l_i^*(j) = k, l_i^*(j) = l_i(j))}{\sum_{j=1}^{n} I(l_i(j) = k)} \tag{15}$$

  The higher is the coverage the better is the classifier.

- **F-measure**: often there is a trade-off between the precision and recall of a classifier. For example, it is easy to make $\mathrm{Recall}_k = 1$ by predicting all testing points to be in class k. However, in this case precision (on k) will be low. On the other hand, we can make precision very high by predicting only a few points as k, for instance, for those predictions where classifier i has the most confidence, but in this case Recall_k will be low. Ideally, we would like both precision and recall to be high. The class-specific F-measure tries to balance the precision and recall values, by computing their harmonic mean for class k

$$F_{\beta,k} = (1+\beta^2)\frac{P_k \cdot R_k}{(\beta^2 \cdot P_k) + R_k} \tag{16}$$

In particular, usually $\beta = 1$:

$$F_k = \frac{2 \cdot P_k \cdot R_k}{P_k + R_k} \tag{17}$$

The higher the F_k value the better the classifier. The overall F-measure for the classifier i is the mean of the class-specific values

$$F = \frac{1}{|K|}\sum_{k=1}^{|K|} F_k \tag{18}$$

For a "perfect" classifier, the maximum value of the F-measure is 1.

For further details, see (Zaki and Meira Jr, 2014).

3.4 JOINT ASPECT-SENTIMENT EXTRACTION

Social tools such as "hashtags" (words or unspaced phrases prefixed with the symbol '#') allow users on Social Networks to easily specify topics they are talking about. Encouraged by this, most of the works in SA (Wang and Manning, 2012; Maas et al., 2011; Go et al., 2009; Barbosa and Feng, 2010) are topic-dependent, i. e. they identify the sentiment of documents given a particular topic filtered a priori. However, this is not always sufficient, because hashtags usually identify the overall topic of a text message and not its sub-topics (or aspects). For example, consider the tweet "*#iOS7 is very good, but they still need to work on battery life and security issues*". The above works would assign the tweet to the topic 'iOS7' because of the hashtag. But the company Apple would be probably more interested in opinions about the aspects '*battery life*' and '*security*' in order to precisely understand how iOS7 could be improved. Moreover, Social Networks contain highly diverse and often latest topics, condensed in short messages that make classical text analysis highly inadequate. Aspect-based sentiment extraction aims to discover different aspects that may occur in a collection of documents. To this purpose, two sub-tasks are needed: (1) extracting all aspect terms (e. g., 'screen') from the corpus, and (2) clustering aspect terms with similar meanings (e. g., cluster 'screen' and 'display' into one aspect category in the domain 'iPhone'). Topic modeling techniques have been successfully applied to tackle this problem. Traditionally, *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003a) and *Probabilistic Latent Semantic Analysis (PLSA)* (Hofmann, 1999) have been used to extract latent aspects from document collections. Once aspects are extracted, the goal is to apply SA techniques on documents focused on the specific aspects. Several works which jointly perform sentiment classification and topic modeling have been proposed. The main advantage of the joint modeling of sentiments and aspects comes from its ability to reciprocally reduce the noise of both tasks. *Topic Sentiment Mixture model (TSM)* (Mei et al., 2007) separates topic and sentiment words using an extended PLSA model. Further models based on the LDA principle can be found in (Lin and He, 2009) and (Jo and Oh, 2011), where the *Joint Sentiment/Topic model (JST)* model and the *Aspect and Sentiment Unification Model (ASUM)*, have been proposed respectively.

Sentiment Analysis has been extensively studied in recent years (Pang and Lee, 2008; Liu, 2012). However, most of the works in SA (Go et al., 2009; Barbosa and Feng, 2010) merely take into account only textual information. They do not consider the relational information embedded in Social Networks. To this purpose, (Speriosu et al., 2011) proposed to enrich the content representation by including the number of followers of the author as additional features. (Tan et al., 2011; Hu et al., 2013) estimated user-level sentiment by exploring tweet contents and friendship relationships among users. These works only address the classification task, but do not extract aspects.

Regarding topic models applied to Social Media, in (Hong and Davison, 2010) standard topic models are used by studying how they can be trained on the dataset. In (Ramage et al., 2010) a scalable implementation of a partially supervised learning model (Labeled LDA) to Twitter feed is presented. However, these models do not address the sentiment classification task and do not include any network information in modeling. (Mei et al., 2008b) regularized a statistical topic model with a harmonic regularizer based on the network structure. In (Chang and Blei, 2010; Sun et al., 2009; Chang and Blei, 2009), the authors proposed relational topic models for document networks that are able to consider both text and structure information. However, these works do not consider sentiment classification. The Multi-Aspect Sentiment model (MAS) (Titov and Mcdonald, 2008) is able to discover aspects from reviews supporting each of the related ratings provided by users. In this work, sentiment (i. e. ratings) is given a priori. The MaxEnt-LDA hybrid model (Zhao et al., 2010) discovers aspects and opinions by incorporating a supervised maximum entropy classifier into an unsupervised topic model. Additional work includes the Topic Sentiment Mixture model (TSM) (Mei et al., 2007) and the Joint Sentiment/Topic model (JST) (Lin and He, 2009). However, all these models do not consider the network information.

3.4.1 *Aspect and Sentiment Unification Model (ASUM)*

This model has been proposed by (Jo and Oh, 2011) and it tackles sentiment classification and aspect extraction at the same time with an unsupervised generative model, discovering pairs of {aspect, sentiment}, which the authors referred to as senti-aspect. ASUM models the generative process of a message as illustrated as follows:

Generative Process: A reviewer first decides to write a review of a restaurant that expresses a distribution of sentiments, for example, 70% satisfied and 30% unsatisfied. And he decides the distribution of the aspects for each sentiment, say 50% about the service, 25% about the food quality, and 25% about the price for the positive sentiment. Then he assigns to each sentence a sentiment and an aspect. For example, he writes that he is satisfied with the friendly service of the restaurant.

Formally, the generative process is as follows (we follow the notations in Table 2):

1. For every pair of sentiment s and aspect z, choose a word distribution $\phi_{sz} \sim \mathtt{Dirichlet}(\beta_s)$.

2. For each message m,
   a) Choose the message's sentiment distribution $\pi_m \sim \mathtt{Dirichlet}(\gamma)$,
   b) For each sentiment s, draw an aspect distribution $\theta_{ms} \sim \mathtt{Dirichlet}(\alpha)$,
   c) For each sentence,
      i. Choose a sentiment $l \sim \mathtt{Multinomial}(\pi_m)$
      ii. Given sentiment l, choose an aspect $k \sim \mathtt{Multinomial}(\theta_{ml})$
      iii. Generate words $w \sim \mathtt{Multinomial}(\phi_{lk})$

Since some positive words, as for example 'good' and 'great', are less probable in negative expressions, and analogously, negative words as 'bad' and 'annoying' are unlikely in positive expressions, this information estimated a priori has been encoded into an asymmetric hyper-parameter β, such that its entries corresponding to general positive sentiment words having small values for negative senti-aspects, and vice-versa. Similarly, the hyper-parameters α and γ represent the information a priori estimated for the aspect distribution θ and the sentiment distribution π in a message m, respectively.

Inference: the latent variables θ, π and ϕ are inferred by Gibbs sampling. At each transition step of the Markov chain, the sentiment

ASUM and **JST**	
T, S, W	number of topics, sentiments and vocabulary size
m, z, s, w	message, aspect, sentiment and word
π, θ, φ	multinomial distribution over sentiments, aspects and words
ASUM	
α_k, γ_l	Dirichlet vectors for θ and π
β_w, $\beta_{(l)w}$	Dirichlet vectors for φ (for sentiment l)
z_{-i}, s_{-i}	vector of assignments of aspects and sentiments for all the sentences in the corpus, except for the i-th sentence
N_{ml}^{-i}	the number of sentences that are assigned sentiment l in message m, except for sentence i
N_{mlk}^{-i}	the number of sentences that are assigned sentiment l and aspect k in message m, except for sentence i
N_{lkw}^{-i}	the number of words that are assigned sentiment l and aspect k, except for sentence i
$m_{i(w)}$	the number of total words (or word w) in sentence i
JST	
α, β, γ	Dirichlet vectors for θ, φ and π
z_{-t}, s_{-t}	vector of assignments of aspects and sentiments for all the words in the corpus, except for the t-th word
$N_{w_t,k,l}^{-t}$	number of times that word w_t appeared in aspect k with sentiment l, except for the t-th word
$N_{k,l}^{-t}$	number of times that words are assigned sentiment l and aspect k, except for the t-th word
$N_{k,l,m}^{-t}$	number of times that word w_t is assigned sentiment l and aspect k in message m, except for the t-th word
$N_{l,m}^{-t}$	number of times that sentiment l is assigned to message m, except for the t-th word
N_{m}^{-t}	total number of words in the corpus, except for the t-th word

Table 2: Meaning of the ASUM and JST notation

and aspect of the i-th sentence are chosen according to the conditional probability

$$p(z_i = k, s_i = l \mid \mathbf{w}, z_{-i}, s_{-i}, \alpha, \beta, \gamma) \propto \frac{N_{ml}^{-i} + \gamma_j}{\sum_{l'=1}^{S} N_{ml'}^{-i} + \gamma_{l'}}$$
$$\frac{N_{mlk}^{-i} + \alpha_k}{\sum_{k'=1}^{T} N_{mlk'}^{-i} + \alpha_{k'}} \frac{\Gamma(\sum_{w=1}^{W} N_{lkw}^{-i} + \beta_{lw})}{\Gamma(\sum_{w=1}^{W} (N_{lkw}^{-i} + \beta_{lw}) + m_i)} \tag{19}$$
$$\prod_{w=1}^{W} \frac{\Gamma(N_{lkw}^{-i} + \beta_{lw} + m_{iw})}{\Gamma(N_{lkw}^{-i} + \beta_{lw})}$$

ASUM is an extension of LDA by restricting the words in a sentence to be drawn from a **single** senti-aspect (i.e. given a sentence, all its words must have the same polarity orientation). If this assumption does not hold, the model does not classify the sentence and simply discards it (e.g., "At the beginning I **hated** *(negative)* iOS7, but then I started to **love** *(positive)* it"). This is an impactful consequence if a message is formed by few sentences. In the original work of (Jo and Oh, 2011), the authors used two lexicons (*PARADIGM* and *PARADIGM+*) which are composed of 7 and 26 seed words, respectively.

Another issue with ASUM is that (Jo and Oh, 2011) only takes into account positive and negative sentiments. In practice, neutral sentiments are the majority and thus need to be distinguished. In this case, the probability of having seed words with different sentiments in the same sentence is even higher. This is because the frequency of neutral words is much higher compared to positive and negative words (e.g., "*The new Apple's OS offers the possibility to block numbers, which is awesome!*" is positive but is composed of many more neutral words than positive words). ASUM used high-level lexicons in (Jo and Oh, 2011) without considering word property in the sentence such as Part-Of-Speech. For example, consider the sentence "*The battery life of iOS7 is like the Android's one*", and suppose we have the word '*like*' in a high-level lexicon as a positive word and all the others having unknown sentiments (non-seed). According to the ASUM's assumption, the non-seed words such as '*battery*', '*life*' and '*Android*' will be treated as positive because the sentiments of words in the same sentence should not have conflicts. Therefore ASUM will classify the overall sentence as positive.

3.4.2 *Joint Sentiment/Topic (JST) model*

In this section, we introduce the *Joint Sentiment/Topic model (JST)* proposed in (He et al., 2011), which is an extension of LDA. LDA framework has three layers (message, topic and word), where topics are associated with messages, and words are associated with topics. In order to model sentiments in messages, a further sentiment layer is added in JST between message and topic layers. The plate notation of JST is very similar to that of ASUM. One of the main difference is that ASUM adds a further layer for sentences, restricting all the words in a sentence to be drawn from a single {aspect, sentiment} pair.

Generative Process: compared with LDA, in which only one message-topic distribution is chosen for one message, S message-topic distributions are sampled. Finally, we draw a word from distribution over words defined by both topic and sentiment labels.

Formally, the generative process is shown as follows (we follow the notations in Table 2):

1. For each message m
   a) Choose a multinomial distribution $\pi_m \sim \mathrm{Dirichlet}(\gamma)$ over sentiment labels.
   b) For each sentiment s under message m, choose a multinomial distribution $\theta_{m,s} \sim \mathrm{Dirichlet}(\alpha)$.
   c) For each word w in message m:
      i. Choose a sentiment label $l \sim \pi_m$,
      ii. Choose a topic $k \sim \theta_{m,s}$,
      iii. Choose a word w from the distribution over words defined by the topic z and sentiment s, φ_z^s, which itself is chosen $\varphi_z^s \sim \mathrm{Dirichlet}(\beta)$.

Inference: the sampling distribution for a word given the remaining topics and sentiment labels can be written as $p(z_t = k, s_t = l \mid \mathbf{w}, z_{-t}, s_{-t}, \alpha, \beta, \gamma)$, where z_{-t} and s_{-t} are vector of assignments of topics and labels for all the words in the corpus except for the t-th word in message m. This equation is derived from marginalizing out the random variables φ, θ and π by Gibbs sampling:

$$p(z_t = k, s_t = l \mid \mathbf{w}, z_{-t}, s_{-t}, \alpha, \beta, \gamma) \propto \frac{N_{w_t,k,l}^{-t} + \beta}{N_{k,l}^{-t} + W\beta} \frac{N_{k,l,m}^{-t} + \alpha}{N_{l,m}^{-t} + T\alpha} \frac{N_{l,m}^{-t} + \gamma}{N_m^{-t} + S\gamma} \tag{20}$$

3.4.3 *Topic Sentiment Mixture (TSM) model*

Topic Sentiment Mixture model (TSM) model (Mei et al., 2007) separates topic and sentiment words using an extended PLSA model. The generative process of TSM is described as follows. An author would "write" an article by making the following decisions stochastically:

1. The author would first decide whether the word will be a common English word (e.g., "the", "a", "of"). If so, the word would be sampled according to the background model θ_B.

2. If not, the author would then decide which of the k subtopics $\theta_1, \ldots, \theta_k$ should be described by the word. A topic model θ in a text collection $\mathcal{C} = \{m_1, \ldots, m_n\}$ is a probabilistic distribution of words $\{p(w \mid \theta)\}_{w \in W}$ and represents a semantically coherent topic. Intuitively, the words of a topic model with high probability often capture the theme of the topic. For example, a topic about "iPhone 6" may assign a high probability to words like "Apple", "battery", "screen", etc.

3. Once the author decides which topic the word is about, the author will further decide whether the word is used to describe the topic neutrally (F), positively (P), or negatively (N).

4. Let the topic picked in step (2) be the k-th topic θ_k. The author would finally sample a word using θ_k (neutral sentiment), the positive sentiment model θ_P or the negative sentiment model θ_N according to the decision in step (3). A sentiment model in a text collection $\mathcal{C}$ is a probabilistic distribution of words representing either positive opinions ($\{p(w \mid \theta_P)\}_{w \in W}$) or negative opinions ($\{p(w \mid \theta_N)\}_{w \in W}$). Sentiment models are orthogonal to topic models in the sense that they would assign high probabilities to general words that are frequently used to express sentiment polarities whereas topical models would assign high probabilities to words representing topical contents with neutral opinions.

The log likelihood of the whole collection C according to the TSM model is

$$\log(\mathcal{C}) = \sum_{m \in \mathcal{C}} \sum_{w \in V} N_{wm} \log[\lambda_B p(w \mid B) + (1 - \lambda_B) \sum_{k=1}^{T} \pi_{mk} \times \\ (\delta_{k,m,F} p(w \mid \theta_k) + \delta_{k,m,P} p(w \mid \theta_P) + \delta_{k,m,N} p(w \mid \theta_N))] \tag{21}$$

where N_{wm} is the count of word w in message m, λ_B is the probability of choosing θ_B (i.e. it indicates how much noise we believe exists in the message collection), π_{mk} is the probability of choosing the topic k in message m, and $\{\delta_{k,m,F}, \delta_{k,m,P}, \delta_{k,m,N}\}$ is the sentiment coverage of topic k in message m. A sentiment coverage of a topic in a message (or a collection of messages) is the relative coverage of the neutral, positive, and negative opinions about the topic in the message (or the collection of messages).

This model is regularized by fixing some parameters. The background model is then set as

$$p(w \mid \theta_B) = \frac{\sum_{m \in \mathcal{C}} N_{wm}}{\sum_{w \in V} \sum_{m \in \mathcal{C}} N_{wm}} \tag{22}$$

The parameters remaining to be estimated are: (1) the topic models, $\theta_1, \ldots, \theta_T$; (2) the sentiment models, θ_P and θ_N; (3) the message topic probabilities π_{mk} ; and (4) the sentiment coverage for each message, $\{\delta_{k,m,F}, \delta_{k,m,P}, \delta_{k,m,N}\}$. Without any prior knowledge, the maximum likelihood estimator is used to estimate all the parameters. Specifically, the Expectation-Maximization (EM) algorithm is used (Dempster et al., 1977) to compute the maximum likelihood estimate iteratively. Further details in (Mei et al., 2007).

4 APPROVAL NETWORK

"You are what you share."
— C. W. Leadbeater,
We Think: The Power
Of Mass Creativity

People who are close to each other in a social network are similar in many ways: they share characteristics and/or act in similar ways. But do users in social networks act similarly because they are close in the network, due to some form of influence that acts along network ties, or rather, are they close in the network because of these similarities, through the processes known as homophily (McPherson et al., 2001)?

By borrowing a nice example from (Shalizi and Thomas, 2011), suppose that there are two friends named Paul and John, and Paul parents ask him the classic hypothetical of social influence: "If your friend John jumped off a bridge, would you jump too?" Why Paul might answer "yes"?

1. because John example inspired Paul (*social contagion/influence*);
2. because John and Paul are friends on account of their shared fondness for jumping off bridges (*manifest homophily*, on the characteristic of interest);
3. because John and Paul became friends through a thrill-seeking club, whose membership rolls are publicly available (*secondary homophily*, on a different yet observed characteristic);
4. because John and Paul became friends through their shared fondness for roller-coasters, which was caused by their common thrill-seeking propensity, which also leads them to jump off bridges (*latent homophily*, on an unobserved characteristic);
5. because John and Paul both happen to be on the Tacoma Narrows Bridge in November 1940, and jumping is safer than staying on a bridge that is tearing itself apart (common *external causation*).

Regarding sentiment classification in social networks, most of the works usually do not take into account relationships among entities (users, posts, etc), and consequently the important information given by the social contagion is left aside. For this reason, the combination of content and relationships is a core task of the recent literature on Sentiment Analysis, where friendships are usually investigated to model the principle of homophily. However, paired with the assumption of homophily, constructuralism explains how social relationships evolve via dynamic and continuous interactions as the knowledge and behavior that two actors share increase. Considering the similarity among users on the basis of constructuralism appears to be a much more powerful force than interpersonal influence within the friendship network. For this reason, a different modeling approach for sentiment analysis is needed. In this chapter Approval Network is presented. It is a novel graph representation to jointly model homophily and constructuralism, which is intended to better represent the contagion on social networks.

4.1 HOMOPHILY AND CONSTRUCTURALISM: TWO SOCIOLOGICAL PROCESSES

Homophily states that a contact among similar people occurs at a higher rate than among dissimilar people

People with different characteristics (e. g., genders, races, ages, class backgrounds, etc.) usually show to have very different personalities: educated people are usually more tolerant, women are usually more sensitive, and gang members are violent (McPherson et al., 2001). Since people generally have significant contacts with others who tend to be like themselves, any personal characteristic tends to converge. **Homophily** is the principle stating that *a contact among similar people occurs at a higher rate than among dissimilar people*. Homophily implies that differences in terms of social characteristics translates into network distance, i. e. the number of relationships through which a piece of information must travel to connect two individuals (McPherson et al., 2001).

The concept of homophily is very ancient. In Aristotle's *Rhetoric* and *Nichomachean Ethics*, he noted that people "love those who are like themselves" (Aristotle, 1934). Plato observed in *Phaedrus* that "similarity begets friendship" (Plato, 1968). However, social scientists began systematic observations of group formation and network ties only in the 1920s (Bott, 1928; Wellman, 1929; Hubbard, 1929). They noted that school children formed friendships and play groups at higher rates if they were similar on demographic characteristics. The classic and most famous work in sociology is (Lazarsfeld and Merton, 1954), where the friendship process is studied. They also quoted the proverbial expression of homophily, "*birds of a feather flock together*", which is often used to summarize this sociological process. Researchers have studied homophily ranging from the strong relationships of "discussing important matters" (Marsden, 1987, 1988) and friendship (Verbrugge, 1977, 1983) to the more circumscribed relationships of "knowing about" someone (Hampton and Wellman, 2000) or appearing with them in a public place (McPherson et al., 1995). In particular, (Lazarsfeld and Merton, 1954) distinguished two types of homophily: **status homophily**, in which similarity is based on informal, formal, or ascribed status, and **value homophily**, which is based on values, attitudes, and beliefs (McPherson et al., 2001). Status homophily includes the major sociodemographic dimensions like race, ethnicity, sex, or age, and acquired characteristics like religion, education, occupation, or behavior patterns. Value homophily includes the wide array of internal states presumed to shape our orientation toward future behavior: attitude, belief, and value similarity lead to attraction and interaction (Huston and Levinger, 1978). Value homophily is the homophily facet that has been considered as assumption in this thesis, where interactions are preferred compared to static user attributes.

Besides homophily, (Carley, 1991) has developed a sociological approach called **constructuralism**, whose core is the assumption that *people who share knowledge each other are more likely to interact* (i. e. form ties). In particular, constructuralism argues that individual learning from interactions takes place on two levels. First, social interactions allow us to collect over time new knowledge that represents bits of larger similarity among users than static sociodemographic dimensions like race, ethnicity, sex, or age (i. e. status homophily). Second, as humans receive and share knowledge with interaction partners, we "learn" a perception of what we expect them to know. Paired with the assumption of homophily that people tend to interact with others similar to them, constructuralism explains how social relationships evolve via interactions as the knowledge that two actors share increases (Joseph et al., 2013). This approach to the coevolution of knowledge and social relationships has considerable explanatory power over the dynamics of social networks and has proved to be an effective tool for social simulation (Carley and Hill, 2001; Hirshman et al., 2011). Since text does not always provide explicit or sufficient information about sentiment, early studies on sentiment classification (Tan et al., 2011; Hu et al., 2013) overcome this limitation by exploiting the principle of homophily, which is usually modeled through friendships. However, considering the similarity among users on the basis of constructuralism appears to be a much more powerful force than interpersonal influence within the friendship network (Kandel, 1978; Cohen, 1977). In other words, considering friendship connections is a weak assumption for modeling homophily mainly for three reasons: (1) being friends does not necessarily mean agreeing on a particular topic (e. g., there are often opposite political views among friends), (2) dynamic interactions are preferred compared to static attributes (value homophily): once friendship is established in online social networks, it is rare that it could be interrupted, or even when it occurs it changes slowly over time, (3) social interactions allow us to collect over time new knowledge that represents bits of larger similarity among users than static sociodemographic dimensions, such as friendship (constructuralism).

Constructuralism states that people who share knowledge with one another are more likely to interact

Considering the similarity among users on the basis of constructuralism appears to be a much more powerful force than interpersonal influence

For these reasons, a different paradigm to jointly model homophily and constructuralism, called **Approval Network** (Pozzi et al., 2013c), is proposed in this chapter (Section 4.4). It is constructed using approval relations among users. For instance, information can spread in Twitter in the form of *retweets*, which are tweets that have been forwarded by a user to his or her followers. A retweet is identified by the pattern "RT @" followed by the name of the tweet's author and the original tweet (e. g., John tweets "*I like the new iPhone*" and Mary retweets the John's tweet: "*RT @John: I like the new iPhone*", i. e. John and Mary positively agrees about iPhone). The corresponding approval tool on Facebook is the "Like" tool. While approving can simply be seen as the act of copying and rebroadcasting, the practice contributes to a conversational ecology in which conversations are composed of a public interplay of voices that give rise to an emotional sense of shared conversational context (Boyd et al., 2010). The general idea behind Approval Network is that a user who approves (e. g., by 'likes' in Facebook or 'retweets' in Twitter) a given message is likely to hold the same opinion of the author. This is because an approval tool does not allow the user to add a comment to the original message to argue against the original post[1]. Thus, "approving" usually means agreeing with the original user: the more are the approvals between two users upon a particular topic of interest, the higher is their agreement on that topic.

The general idea behind Approval Network is that a user who approves a given message is likely to hold the same opinion of the author

4.2 RELATIONAL DATA REPRESENTATION

Relational data is formed of a complex structure that can be represented in a multitude of ways. Due to this, in order to understand which is the best relational data representation for the Approval Network, we need to consider that the choice of a specific data representation is likely to impact both on the applicability of particular models and their performance. When choosing data representation in a relational context, there are two categories of decisions that can be considered. First, relational data can be propositionalized for the application of non-relational learning algorithms by performing a merely "flattening" approach, producing a misleading and poor data representation. A **propositional representation** of real world data leads us to a substantial loss of information, provided by the relational data structure, and to the introduction of a statistical bias. Once data is propositionalized, traditional model learning can be subsequently performed. However, in order to fully exploit the relational information, SRL researchers have chosen to represent the data using attributed graphs (see e. g., (Friedman et al., 1999)). For this reason, this thesis focuses on the **graph-based representation**, which has been a common and

A propositional representation of real world data leads us to a substantial loss of information

1 Note that for this reason the Facebook's '*Share*' tool does not belong to approval tools.

powerful choice for addressing the growing interest in online social networks such as Facebook, Twitter, Flickr, and LinkedIn (Mislove et al., 2007; Ahmed et al., 2010). Specifically, a graph-based data representation is defined as $G = \langle V, E, \mathbf{X}^V, \mathbf{X}^E \rangle$ where the nodes V are entities and the links E represent relationships among those entities. $\mathbf{X}^V$ is a set of features about the entities in V, and $\mathbf{X}^E$ the set of features that provides information about the relation links in E. As an example, we consider hypothetical data inspired by Facebook. Nodes V can be users and the links E the friendship relations among them. $\mathbf{X}^V$ is a set of features about the users in V such as their gender, location, relationship status, school, favorite movies, or musical preference. Likewise, the set of features $\mathbf{X}^E$ provides information about the friendship links in E such as the time of formation or the number of messages exchanged between two users. The hypothetical goal could be the prediction of the political party (Democratic or Republican) of every node (user) in G, assuming that this affiliation is known for some but not all of the users in G (i. e. a semi-supervised approach). Traditional classification approaches (based on the i.i.d. assumption) ignore the friendships (i. e. links) and classify the users using only gender, location or other user information. However, the inclusion of relational information in the classification task could potentially increase the classification performance. For instance, a feature could be computed considering, for each user, the proportion of friends that are male or that live in a particular region.
Conversely, a further aim could also be the estimation of user features given their political affiliation (Jensen et al., 2004). For instance, the users' gender and location could be estimated starting from their political affiliations. Many such algorithms have been proposed in the literature, such as *Gibbs Sampling* (Jensen et al., 2004), *relaxation labeling* (Chakrabarti et al., 1998), *belief propagation*(Taskar et al., 2002), *ICA* (Neville and Jensen, 2000; Lu and Getoor, 2003), and *weighted neighbor techniques* (Macskassy and Provost, 2007).

In both of the directions for data representation (propositionalized data and graph-based representation) an important role is played by the relationships among instances, not only in terms of their presence/absence but also in terms of probabilistic connections. For example, if two Facebook users have opposite probabilities on the edges to belong to a specific political party (e. g., Democratic or Republican), their likelihood to belong to the same political party needs to be decreased. On the other side, if two users are both characterized by a high probability edge, their likelihood to belong to the same party should be increased. The presence of probabilistic relationships offers a unique and valuable opportunity to improve model performance, either by creating a dataset representation that takes into account probabilistic connections during the propositionalization activities, and by inducing inference models driven by features, instances, and uncertainty over relationships.

Once the type of representation has been defined, the specific content of the data representation need to be considered. For instance, features for the nodes and links of a graph can be constructed and populated using a wide range of aggregation functions (e. g., a link between two users can be weighted by averaging their respective number of friends). In the following, a detailed review of the existing techniques for graph transformation are presented.

4.3 GRAPH TRANSFORMATION

This section is aimed at examining and categorizing various techniques for transforming the representation of graph-based relational data. These changes are typically seen as a pre-processing step that increases performance and/or speed for some other subsequent task, such as object classification. However, an output of these techniques can itself be valuable (e. g., the "*People you may know*" service, proposed in several social networks, is a link prediction transformation).

Given a set of graph-based relational data, relational representation transformation is defined as any change to the space of links, nodes, and/or features. For instance, in Figure 6 the original graph

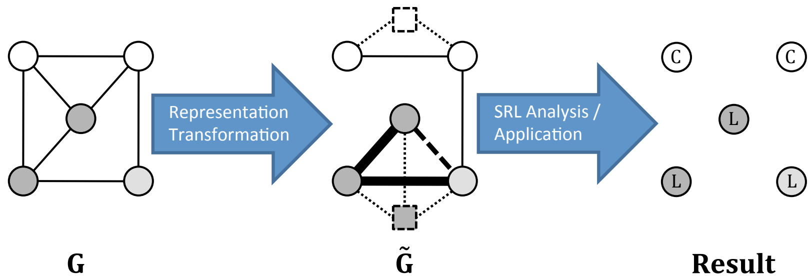


Figure 6: Example Transformation and Subsequent Analysis

representation G is transformed into a new representation $\tilde{G}$, where links, nodes, and features have been added, removed or modified. Some SRL algorithm or analysis can subsequently be applied to $\tilde{G}$ (e. g., node classification). The particular transformations will vary depending on the desired application. For instance, (Gallagher et al., 2008) found that adding links among similar nodes increases node classification accuracy up to 15% on some tasks. Similarly, (Neville and Jensen, 2005) showed that adding nodes representing underlying groups leads to both simpler inference and a higher accuracy.

Figure 7 shows a taxonomy proposed in (Rossi et al., 2012) regarding the relational representation transformation tasks. The two main tasks in this taxonomy are **link transformation** and **node transformation**, which can be subsequently decomposed into *prediction* and *interpretation* tasks. The former task involves predicting the existence

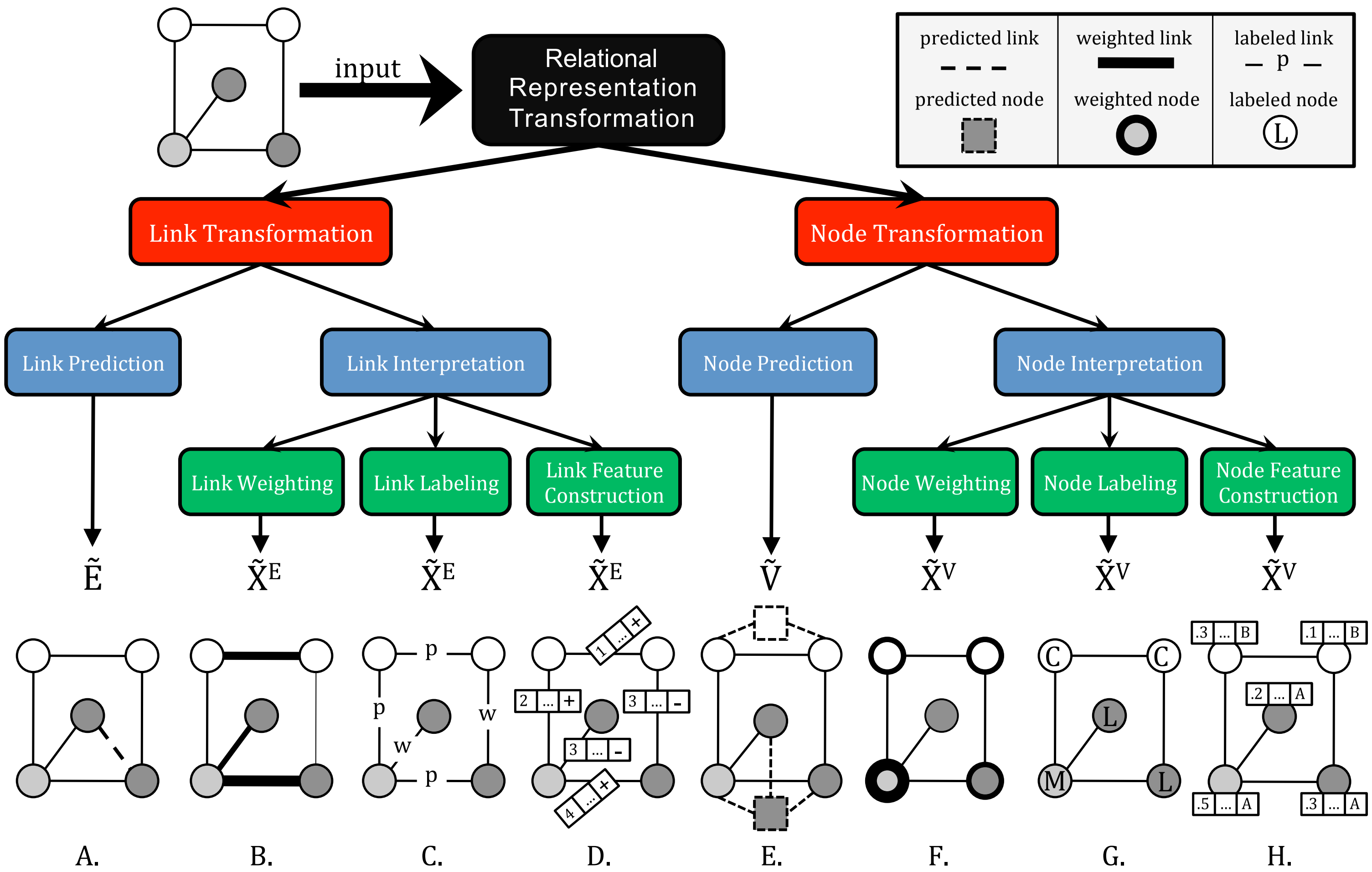


Figure 7: Relational Representation Transformation Taxonomy: Link and node transformation are formulated as link prediction, link interpretation, node prediction, and node interpretation tasks. Interpretation is further divided into weighting, labeling, or constructing features.

of new nodes and links, while interpretation involves the weights, labels, or features construction of nodes/links.

First, **link prediction** adds new links to the graph. The sample graph A in Figure 7 shows a link being predicted based on the similarity between two nodes. There are many simple link prediction algorithms based on similarity, neighbor properties, shortest path distances, infinite sums over paths (i. e. random walks), and other strategies (Rossi et al., 2012).

Second, there are several types of **link interpretation**, which involves constructing weights, labels, or features for the existing links. For instance, in many graphs (including Facebook data), not all links (or friendships) have equal importance. Sample graph B in Figure 7 shows the result of performing **link weighting**. Weights are usually estimated based on the assumption that high similarity (among feature values of each pair of linked nodes) may indicate stronger relationships. Alternatively, **link labeling** is used to assign some kind of label to links. For instance, based on known feature values between the linked users, links might be labeled as either "personal" (p) or "work" (w) (sample graph C of Figure 7). Finally, graph D of Figure 7 shows how **link feature construction** can be used to add feature values to each link. For instance, a link feature might count the number of common friends.

Third, **node prediction** adds additional nodes to the graph[2]. For instance, graph E of Figure 7 shows the two latent groups discovered through relational clustering. Node prediction techniques can be used to identify people who are particularly similar based on their political affiliation.

Similarly to links, there are several types of **node interpretation**, which involves constructing weights, labels, or feature values for existing nodes. For instance, some nodes may be more influential/important than others and thus should have a higher weight. Sample graph F of Figure 7 shows node weighting, where the weights might be assigned based on the numbers of friends or via the PageRank/eigenvector techniques (more generally, centrality measures). Node weighting techniques are typically applied on social media for opinion leader detection tasks. Sample graph G of Figure 7 shows an example of **node labeling**. In the example, a label representing conservative (C), liberal (L), or moderate (M) has been assigned to each node. Such labels can be estimated (1) using the non-relational features, (2) using relations (3) via textual analysis, (4) a mixture of (1), (2) and (3). Finally, sample graph H of Figure 7 shows the result of **node feature construction**, where feature values are added to nodes. Any feature that is correlated with political affiliation could be used (and stored in the graph-based representation) to improve the performance of a classification algorithm. For instance, naïvely suppose we find that users

2 Relative links are also added

with relatively few Facebook friends are often Republican while those with many friends are often Democratic. In this case, a counter of the number of friends would be stored as an additional node feature.

Among the eight transformation tasks reported in the presented taxonomy (Figure 7), only *Link Weighting*, *Node Labeling* and *Feature Construction*, used in Section 4.4 to build the Approval Network, are described in details:

- **Link Weighting**: given the initial graph $G = \langle V, E, \mathbf{X}^V, \mathbf{X}^E \rangle$, the task is to assign a weight (typically continuous) to each existing link in G, representing the importance or influence of that link. Unlike link prediction, most link weighting techniques are designed to work only with links that already exist in the graph. In the simplest case, link weighting can be performed just aggregating a intrinsic properties of links. If link features like duration, direction, or frequency are known, they can be aggregated in some way to generate link weights. For example, (Onnela et al., 2007) defines link weights based on the aggregated duration of phone calls between users in a mobile communication network. If link weights are already known for some of the links, then supervised methods can be used for weight prediction. For instance, (Kahanda and Neville, 2009) predict link strength within a Facebook dataset, where stronger relationships are identified based on a user's explicit identification of their "top friends" via a popular Facebook application. (Gilbert and Karahalios, 2009) generate features based on profile similarity (e. g., if two users have similar education levels) and based on user interactions (e. g., about what topics and how frequently two users communicate).

  Moreover, some researchers have considered the importance of time in evaluating link weight, under the assumption that interactions that occurred more recently should have a stronger weight. For instance, (Roth et al., 2010) propose the "Interactions Rank" metric: it weights incoming and outgoing messages for each link, and imposes an exponential decay on the importance of each message based on how old it is.

- **Node Labeling**: given the initial graph $G = \langle V, E, \mathbf{X}^V, \mathbf{X}^E \rangle$, the task is to assign some labels to nodes in G. In many cases, node labeling may be itself considered a goal (e. g., the prediction of the political affiliation in our example). However, node labeling can also be considered as a representation transformation for the desired task. For instance, for anomalous link detection task, having estimated node labels would allow us to identify links between nodes whose labels indicate they should not be connected (Rattigan and Jensen, 2005).

Even when node labeling is the final goal, intermediate label estimation may still be useful as a representation transformation. In particular, (Kou and Cohen, 2007) describe a "stacked model" for relational classification that relabels the training set with estimated node labels using a non-relational classifier. They then use these estimated labels to learn a new relational classifier.

Node labeling can of course also be done by non-relational supervised classifiers such as SVM, decision trees, kNN, logistic regression, and Naïve Bayes. These methods simply use features $\tilde{X}^V$ and do not exploit the graph topology. However, instead of working on manually tagged datasets, labels can also be automatically assigned via unsupervised techniques. There are many networks in the real-world that contain textual content such as social networks, email/communication networks, citation networks, and many others. Traditional textual analysis models such as Latent Semantic Allocation (LSA) (Deerwester et al., 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) can be used to assign topic representing a summary of the textual information to each node (this task will discussed in Section 3.4). In particular, more recent techniques such as Link-LDA (Erosheva et al., 2004) and Link-PLSA (Cohn and Hofmann, 2001) have been proposed to incorporate links into the traditional techniques. (Cohn and Hofmann, 2001) shows that including links can produce more accurate node labels than techniques that use only the node attributes.

- **Feature construction**: it is the process aimed at forming new features for nodes/links. Feature construction is very frequently done before performing a task such as classification and typically improve the accuracy or understandability of SRL algorithms. A feature can be categorized according to the types of information used in the construction process. The possible information to use includes the set of nodes V or links E, the node features X^V , and the link features X^E. The categorization consists of four types:
  - **Non-relational Features**: a feature is considered a non-relational feature if the value of the feature for a particular node/link is computed using only the non-relational features (i. e. , attributes) of nodes, ignoring any link-based information. A new feature value might be constructed from an existing feature vector by adding together several feature values, by thresholding a particular value, etc.
  - **Topology Features**: a feature is considered a topology-based feature if values of the feature are computed using only the nodes V and/or links E, ignoring any existing node

and link feature values. For instance, the new feature value might count the number of adjacent nodes, or how many shortest paths in the graph pass through the target node.

- **Relational Link-value Features**: a feature is considered a relational link-value feature if the values of the links are used for computing the new feature. Typically, some kind of aggregation operator is applied to these values, such as count, mode, average, proportion, etc. For instance, a new node feature might be computed through the average of link-values that represent the number of messages exchanged with friends.
- **Relational Node-value Features**: a feature is considered a relational node-value feature if the values of nodes are used in the construction. For instance, a new feature might count the number of friend nodes and another might count the number of "People you may know" nodes.

Table 3 shows a comprehensive categorization of the state-of-the-art techniques that can be used when dealing with Graph Transformation (Rossi et al., 2012).

	LINKS	NODES
Prediction	• Adamic/Adar (Adamic and Adar, 2003), Katz (Katz, 1953), and others (Liben-Nowell and Kleinberg, 2003) • Text or Feature Similarity (Macskassy, 2007) • Classification via RMN (Taskar et al., 2003) or SVM (Hasan et al., 2006)	• Spectral Clustering (Neville and Jensen, 2005), Mixed-Membership Relational Clustering (Long et al., 2007) • LDA (Blei et al., 2003b), PLSA (Hofmann, 1999) • Hierarchical Clustering via Edge-betweenness (Newman and Girvan, 2004)
Weighting	• Latent Variable Estimation(Xiang et al., 2010) • Linear Combination of Features (Gilbert and Karahalios, 2009) • Aggregating Intrinsic Information (Onnela et al., 2007)	• Betweenness (Freeman, 1977), Closeness (Sabidussi, 1966) • HITs (Kleinberg, 1999), Prob. HITs (Cohn and Chang, 2000), Sim-Rank (Jeh and Widom, 2002) • PageRank (Page et al., 1999), Topical PageRank (Haveliwala, 2003; Richardson and Domingos, 2002)
Labeling	• LDA (Blei et al., 2003b), PLSA (Hofmann, 1999) • Link Classification via Logistic Regression (Leskovec et al., 2010), Bagged Decision Trees (Kahanda and Neville, 2009)	• LDA (Blei et al., 2003b), PLSA (Hofmann, 1999) • Node Classification via Stacked Model (Kou and Cohen, 2007) or RN (Macskassy and Provost, 2003)
Feature Construction	• Link Feature Similarity (Rossi and Neville, 2010) • Link Aggregations (Kahanda and Neville, 2009) • Graph Features (Lichtenwalter et al., 2010)	• Database Query Search (Popescul et al., 2003), RPT (Neville et al., 2003a) • MLN Structure Learning (Kok and Domingos, 2009, 2010) • FOIL, nFOIL (Landwehr et al., 2005), kFOIL (Landwehr et al., 2010), Aleph (Srinivasan, 1999)

Table 3: Techniques for Relational Representation Transformation

4.4 APPROVAL NETWORK FORMALIZATION

Once the graph transformation techniques existing in the literature have been presented (Section 4.3), te key components that allow us to explicitly use approval relations can now formally defined:

Def. 4.1 *Given a query (i.e. topic) of interest* q*, a* ***Directed Approval Graph*** *is a quadruple* $DAG_q = \{V_q, E_q, \mathbf{X}_q^V, \mathbf{X}_q^E\}$*, where* $V_q = \{v_1, ..., v_n\}$ *represents the set of active users on* q*;* $E_q = \{(v_i, v_j) | v_i, v_j \in V_q\}$ *is the set of approval edges, meaning that the extent that* v_i *approved* v_j*'s messages;* $\mathbf{X}_q^E = \{w_{i,j} | (v_i, v_j) \in E_q\}$ *is the set of weights assigned to approval edges, indicating that* v_i *approved* $w_{i,j}$ *messages of* v_j *on* q*;* $\mathbf{X}_q^V = \{k_i | v_i \in V_q\}$ *is the set of coefficients related to nodes, where* k_i *represents the total number of messages of* v_i *on* q*.*

Given a DAG_q, the Normalized Directed Approval Graph is defined in (Pozzi et al., 2013c) as:

Def. 4.2 *Given an Approval Graph* $DAG_q = \{V_q, E_q, \mathbf{X}_q^V, \mathbf{X}_q^E\}$*, a* ***Normalized Directed Approval Graph*** *is derived as a triple* $N\text{-}DAG_q = \{V_q, E_q, \mathbf{C}_q^E\}$*, where* $\mathbf{C}_q^E = \{c_{i,j} | w_{i,j} \in \mathbf{X}_q^E, k_j \in \mathbf{X}_q^V\}$ *is the set of normalized weights of approval edges, and* $c_{i,j}$ *is calculated as*

$$c_{i,j} = \frac{w_{i,j}}{k_j} \tag{23}$$

In addition to the weighting schema presented in Equation (23), an alternative index is proposed to overcome some limitations. First, the common characteristic of an approval network is that most of the users usually approve only one message of a target user, and very few users approve two or more messages (Kwak et al., 2010). Thus, for a better approximating, a logarithmic distribution should be sued instead of a linear one. Second, the number of approvals between two users does not necessarily indicate how much they agree with a particular topic. It could be influenced by the interest and originality of the target user's messages. For example, a user A could completely agree with user B but approves it one or two times only because the weak originality of B's messages.

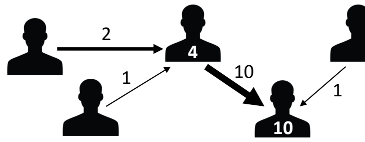
For this reason, the number of approvals from user A to B considering the maximum number of approvals from any user connected to B are normalized. Both aspects converged in the following definition:

Def. 4.3 *Given an Approval Graph* $DAG_q = \{V_q, E_q, \mathbf{X}_q^V, \mathbf{X}_q^E\}$*, an* ***Augmented Directed Approval Graph*** *is derived as a triple* $A\text{-}DAG_q = \{V_q, E_q, \mathbf{C}_q^E\}$*, where* $\mathbf{C}_q^E = \{c_{i,j}\}$ *is the set of normalized weights of approval edges, and* $c_{i,j}$ *is calculated as*

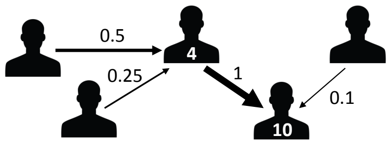
$$c_{i,j} = \frac{w_{i,j}}{\max_i w_{i,j}} \log_2 \left(1 + \frac{w_{i,j}}{k_j}\right) \tag{24}$$

Note that DAG_q, $N\text{-}DAG_q$ and $A - DAG_q$ are formed through a link weighting task in graph transformation, discussed in Section 4.3.

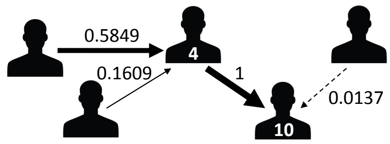
Figure 8 shows the differences between Normalized and Augmented DAG_q (Def. 4.2 and Def. 4.3).



(a) Network example



(b) Simple index (Def. 4.2)



(c) Augmented index (Def. 4.3)

Figure 8: Examples of Normalized and Augmented DAG_q modeling approval relations. White numbers represent the quantity of messages emitted by every user. The weight on the edges of (a) is the number of approvals that a source user makes among messages of the target user.

As mentioned, most of the users usually approve only one message of a target user, but this is common practice on Social Networks. But approving only one message does not lead to valuable information regarding the agreement among users. Figure 8(c) shows that Augmented index is able to give a lower weight to edges where the number of approval is 1. Moreover, Augmented index penalizes users who approve few tweets of a particular target user if there are other users who approve many tweets of the same target user (0.0137 in Figure Figure 8(c)).

In order to create a unique representation both for user-user (approval) and user-message relationships, A-DAG_q has been extended to deal with heterogeneous graphs:

Def. 4.4 *Given a A-$\text{DAG}_q = \{V_q, E_q, \mathbf{C}_q^E\}$, let $M_q = \{m_1, \cdots, m_m\}$ be the set of nodes representing messages about q and $A_q^M = \{(v_i, m_t) | v_i \in V_q, m_t \in M_q\}$ be the set of arcs that connect the user v_i and the message m_t. A **Heterogeneous Directed Approval Graph** is a quintuple H-$\text{DAG}_q = \{V_q, E_q, C_q^E, M_q, A_q^M\}$.*

A graphical representation of H-DAG is reported in Figure 9. In the following, topic q is intended to be fixed and therefore omitted and H-DAG will be denoted as ϕ.

In the following, a real approval network is shown in Figure 10. It is based on retweets of users posting about the topic iOS7. It is composed of 30206 users and 24779 connections. This dataset has been used in the work that will be described in Chapter 7, where a model is proposed to simultaneously extract aspects and classify sentiments from textual messages using the network information.

As mentioned, a common practice on Social Networks lies in approving only one message by most of the users. This is shown in

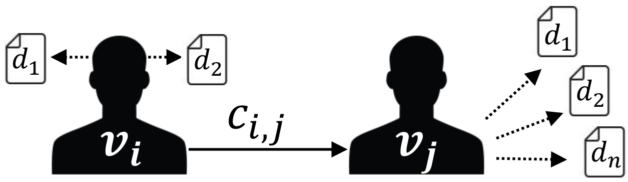
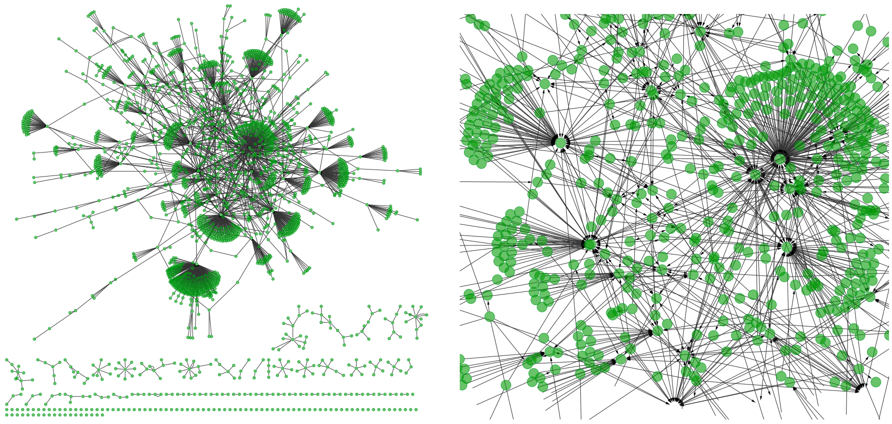


Figure 9: H-DAG representing user-message and user-user (approval) dependencies



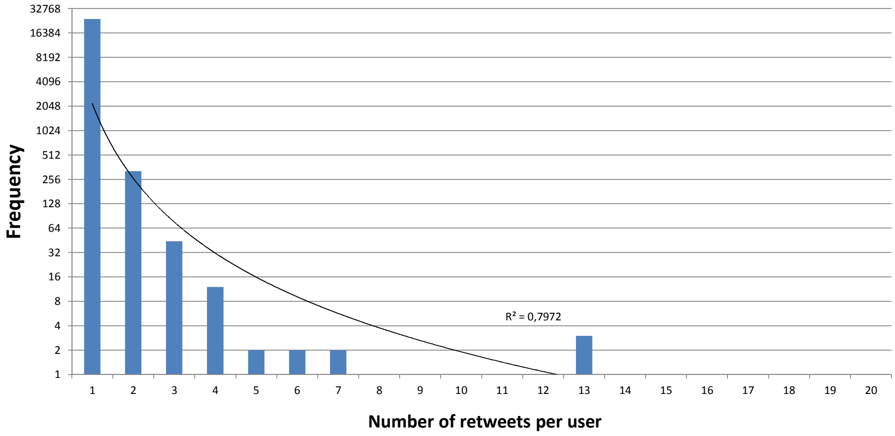
(a) Original view (b) Detailed view

Figure 10: Real approval (sub-) network based on retweets of users posting about the topic iOS7

Figure 11, where the decay in frequency became even more evident according to the increase of the number of retweets per user. The second histogram of Figure 11 shows the normal behavior of the social network users: given a topic of interest, most of them just tweet a single message. Even in this histogram, the decay approximates a logarithmic distribution ($R^2 = 0.894$) according to what stated by the Power-law (Adamic et al., 2001; Clauset et al., 2009).

4.5 PROBABILISTIC RELATIONAL MODELS

Once the Approval Network has been formally defined, the issue related to its inclusion in relational classification models has to be addressed. Traditional Machine Learning methods are no longer effective for mining opinions in relational environments, because the majority of research assumes independently and identically distributed data (i.i.d.). However, the problem of dealing with uncertainty and complex relational data structures (where the i.i.d. assumption is violated) has been addressed and investigated in the last years. The goal of creating efficient representation for complex relational data structures deals with Relational Database, Probabilistic Relational Database and First-Order Knowledge Base, while the goal of managing uncertainty is tackled by Statistical Learning and more in general by



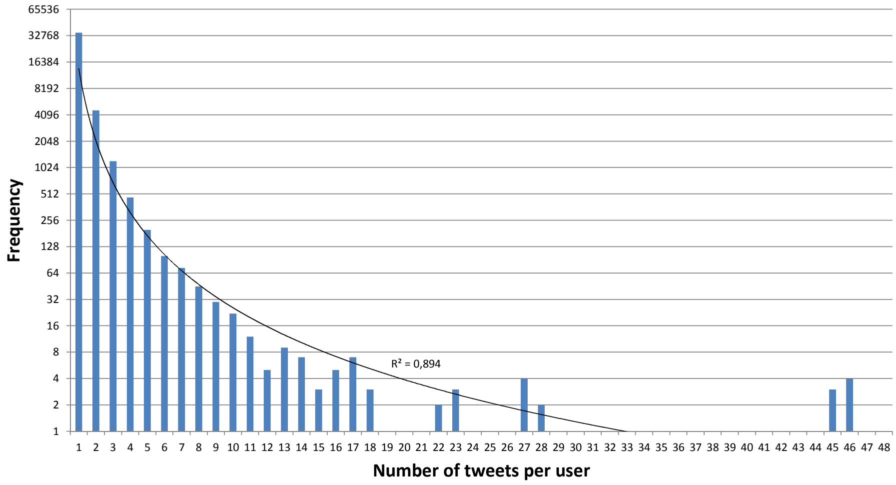


Figure 11: Histograms of the iOS7 network

Data Mining techniques. When relationships among entities are taken into account, Statistical Learning takes the name Statistical Relational Learning (SRL). This thesis is focused on SRL, which can be viewed, as depicted in Figure 12, as the intersection of three main research areas: Relational Representation, Learning Techniques and Probabilistic Models.

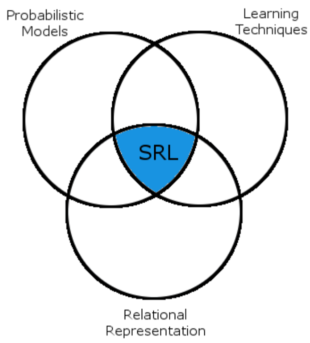


Figure 12: Statistical Relational Learning (SRL)

The notion of model representation is also important in SRL: the choice of what kind of statistical model is suitable to represent the relationships is an important task. Some of the most prominent models for SRL are *Probabilistic Relational Models (PRMs)* (Friedman et al., 1999), *Relational Markov Networks (RMNs)* (Taskar et al., 2002), *Relational Dependency Networks (RDNs)* (Neville and Jensen, 2007), *Structural Logistic Regression* (Popescul et al., 2003), *Conditional Random Fields (CRF)* (Lafferty et al., 2001), and *Markov Logic Networks (MLNs)* (Domingos and Richardson, 2004; Richardson and Domingos, 2006).

However, the choice of the most "appropriate" statistical model does strongly interact with what kinds of node and link features are useful. While a number of relevant comparisons have already been published (Jensen et al., 2004; Neville and Jensen, 2007; Macskassy and Provost, 2007; Sen et al., 2008; Crane and McDowell, 2011), more work is needed to evaluate the interaction between the choice of the statistical model and feature selection, and to evaluate which statistical models work best in domains (Rossi et al., 2012).

Probabilistic Relational Models incorporate both vertex and edge attributes to model the joint probability distribution of a set of entities and the links that associate them

Probabilistic Relational Models (PRMs) are a rich representation language which provide a systematic way to incorporate both vertex and edge attributes to model the joint probability distribution of a set of entities and the links that associate them. The benefit of a PRMs is that it considers the object-relational nature of structured data by capturing probabilistic interactions between entities and the links themselves. From a high level point of view, PRMs specify a probability model over attributes of interrelated objects rather then over features of propositional samples. The simplest form of PRMs was introduced by (Friedman et al., 1999). PRMs are mainly composed of two key components: (1) a relational structure over attributes with the associated parameters; (2) a joint probability model over relational data.

There are two pioneering approach of PRMs, one based on *Bayesian networks*, which consider the links to be directed (Getoor et al., 2003), and the other based on relational *Markov networks*, which consider the relation links to be undirected (Taskar et al., 2003). Moreover, in (Fersini et al., 2009) PRMs have been investigated and extended in order to measure and include uncertainty over relationships.

As an example to abstractly explain how PRMs work, consider the link prediction problem in a Facebook network. The only entity that non-relational models consider is the user and their attributes. However, heterogeneous entities can be mixed in PRMs giving the possibility to include other relevant objects in this model, such as messages. Similar to a database schema, each of these objects can have attributes. For example, a user may have attributes like name, gender, location, etc, while a message may have timestamp and geo-location. Moreover, there can be relational links between these entities. For instance, two users can be connected by a friendship relationship. A user can be also connected to messages by author relationships. In this way, the model can include a complete relational schema similar to an object relational database.

PRMs were originally designed for the attribute prediction problem in relational data (Aggarwal, 2011) and has been subsequently extended for the link prediction task (Getoor et al., 2003; Taskar et al., 2003). In this case, additional objects are added in the relational schema. Any link object l, is associated with a tuple of entity objects $(o_1, o_2, \ldots, o_k)$ that participate in the relation. Following the previous example, one of the link object can be the friendship that relates two users. The model also allows the link objects to have attributes. Now, consider a object named "Suggested Friendship" that relates two users. It could be a binary attribute which is true if there exists a link between the associated objects, and false otherwise. More accurately, it could be a probabilistic attribute that represents the probability of two connected users to be effectively friends. The link prediction task now reduces to the problem of predicting the binary/continuous attribute value of these link objects. In the training step of the model, a single probabilistic model is defined over the entire graph, including both object labels and links between the objects. The model parameters are trained discriminatively to maximize the probability of the object and the link labels given the known attributes. The learned model is then applied using probabilistic inference, to predict and classify links using observed attributes and links. In particular, the works presented in Chapter 6 and Chapter 7 propose particular types of PRMs, where the network information (Approval Network) and the content of messages are combined to predict sentiment classification.

5 A CONTENT-BASED APPROACH FOR POLARITY CLASSIFICATION AT DOCUMENT-LEVEL: BAYESIAN MODEL AVERAGING

"Life should be like a good Tweet - short, pithy, convey a message and inspire others to follow."
— Ashok Kallarakkal

As presented in Section 3.3.4, there are mainly three kinds of analyses in the context of social networks: (1) *Link-based Analysis*, where only structural relationships are used, (2) *Content-based Analysis*, where only contents is leveraged and (3) *Combining Link-based and Content-based Analysis*, particularly useful when textual features do not provide sufficient information to infer sentiment (e. g., *I agree!*). Polarity classification can be approached by Content-based Analysis and combining Link-based and Content-based Analysis, while Link-based analysis is usually used to simulate and analyze the diffusion model of opinions across the network. From this chapter three contributions for polarity classification will be presented. The first contribution is presented in this chapter. It concerns with a content-based methodology based on the ensemble learning mechanism, which is also used in the second contribution (Chapter 6), together with the network structure, for user-level Sentiment Analysis. The results are compared with a methodology that estimates the user polarity aggregating the polarities of its messages estimated through a content-based approach. The third contribution (Chapter 7) proposes an aspect-sentiment modeling with network information to simultaneously extract aspects and classify sentiments from textual messages.

5.1 BACKGROUND

Although there are relevant unsupervised and semi-supervised methods in the literature (Section 3.3), most of the approaches for polarity classification focus on supervised learning, thanks to their predictive power (Liu, 2012). The common characteristic of these approaches concerns with the identification of the model which classify the polarity of text sources with the highest accuracy as possible. However, none of the classification algorithms consistently perform better than others and there is no consensus regarding which methodology should be adopted for a given problem in a given domain. Moreover, the selection of the best single classification model often leads

There is no consensus regarding which classification algorithm should be adopted for a given problem in a given domain to over-confident decisions that do not take into account the inherent uncertainty of the natural language. Therefore, in order to overcome these limitations, an ensemble of different classifiers could lead to more robust and accurate classification (Dietterich, 2002; Bauer and Kohavi, 1999). The idea behind ensemble mechanisms is to exploit the characteristics of several independent learners by combining them in order to achieve better performance than the best baseline classifier. As mentioned in (Windeatt and Ardeshir, 2004), two necessary conditions should be satisfied to achieve a good ensemble: accuracy and prediction diversity. The state of the art about ensemble learning for SA basically comprises traditional methods such as *Majority Voting (MV)*, *Bagging* and *Boosting* (see (Wang et al., 2014) for a comprehensive study). MV is the most widely used ensemble technique, which is characterized by a set of "experts" that classifies the sentence and determines the final polarity by selecting the most popular label prediction to increase the accuracy but not explicitly addressing diversity. Further approaches aimed at accounting for diversity are represented by Bagging and Boosting. In Bagging, diversity is obtained by using bootstrapped replicas of the training data: different training data subsets are randomly drawn, with replacement, from the entire training dataset. Each training data subset (bag) is used to train a different baseline learner of the same type. Regarding *Boosting*, it incrementally builds an ensemble by training each new model to emphasize those instances that previous models misclassified.

Although the presented approaches are widely used in SA, they suffer of several limitations that the proposed method, described in the following, intends to overcome:

- **Single learner generalizes worst than multiple models.** Many machine learning approaches have been investigated for sentiment classification purposes (Pang and Lee, 2008). However, within the sentiment classification research field, there is no agreement on which methodology is better than others: one learner could perform better than others in respect of a given application domain, while a further approach could outperform the others when dealing with a given language or linguistic register. The uncertainty about which model provides the highest performance in different context can be overcome by introducing an ensemble learning approach able to exploit the potentials of several learners when predicting the sentiment orientation.

- **Ensembles assume independent and equally trustworthy models.** Classifiers enclosed in traditional ensemble learning approaches are assumed to be independent and equally reliable (Hassan et al., 2013; Su et al., 2013; Wang et al., 2014), which is not true in a real situation. For instance, consider several poor classifiers which make highly correlated mistakes predicting positive sentences as negative and a good classifier that correctly predicts the sentiment orientation. Assuming these classifiers as independent and equally reliable could lead to biased decisions. An ensemble method could take into account dependencies and accuracies of learners that would help to evaluate the contribution of each model in an ensemble and to smooth weak classifiers when making polarity predictions.

- **The search of the optimal ensemble composition comes with a cost.** One of the major challenges is concerned with online big data, where ensembles are attempting to come up with a reasonable trade-off between classification accuracy and computational time. Traditional state of the art approaches mainly focus on dealing with data and/or models to obtain the highest recognition performance, disregarding the computational complexity issue.

- **Lack of investigations across several domains.** Traditional ensemble approaches have shown their potential in predicting the sentiment orientation either on well-formed texts (e. g., reviews) (Dave et al., 2003; Xia et al., 2011) or on noisy contents (e. g., tweets) (Hassan et al., 2013), but not both.

Starting from the idea proposed in (Pozzi et al., 2013a), Bayesian Model Averaging (BMA) has been proposed to overcome these limitations by a novel Bayesian Ensemble Learning paradigm (Fersini et al., 2014b), where the marginal predictive capability of each model is taken into account. A greedy selection strategy, based on backward elimination, is used to derive the optimal ensemble of classifiers.

5.2 THE CLASSIFICATION MODEL

The most important limitation introduced by Bagging and MV, as mentioned in the previous section, is that the models to be included in the ensemble have uniform distributed weights regardless their reliability. However, the uncertainty left by data and models can be filtered by considering the Bayesian paradigm. In particular, all the possible models in the hypothesis space could be exploited by considering their marginal prediction capabilities and their reliabilities. Given a sentence s and a set C of independent classifiers, the probability of label $l(s)$ is estimated by **Bayesian Model Averaging (BMA)** as follows:

$$P(l(s) \mid C, \mathcal{D}) = \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(i \mid \mathcal{D}) \tag{25}$$

where $P(l(s) \mid i, \mathcal{D})$ is the marginal distribution of the label predicted by classifier i and $P(i \mid \mathcal{D})$ denotes the posterior probability of model i. The posterior $P(i \mid \mathcal{D})$ can be computed as:

$$P(i \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid i)P(i)}{\sum\limits_{j \in C} P(\mathcal{D} \mid j)P(j)} \tag{26}$$

where $P(i)$ is the prior probability of i and $P(\mathcal{D} \mid \cdot)$ is the model likelihood. In Equation (26), $P(i)$ and $\sum_{j \in C} P(\mathcal{D} \mid j)P(j)$ are assumed to be a constant and therefore can be omitted. Therefore, BMA assigns the optimal label $l^*(s)$ to s according to the following decision rule:

$$\begin{aligned} l^*(s) = \arg\max_{l(s)} P(l(s) \mid C, \mathcal{D}) &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D})P(i \mid \mathcal{D}) \\ &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D})P(\mathcal{D} \mid i)P(i) \\ &= \sum_{i \in C} P(l(s) \mid i, \mathcal{D})P(\mathcal{D} \mid i) \end{aligned} \tag{27}$$

The implicit measure $P(\mathcal{D} \mid i)$ can be easily replaced by an explicit estimate, known as F_1-measure (Section 3.3.5), obtained during a preliminary evaluation of the classifiers i. In particular, by performing a cross validation each classifier can produce an averaged measure stating how well a learning machine generalizes to unseen data. Note that the weights of classifier i have been computed on each fold separately, i. e. the weights to be used for inference in a particular fold are estimated on the remaining ones. This procedure has been adopted to avoid over-fitting. Considering ϕ-folds for cross validating a classifier i, the measure $P(\mathcal{D} \mid i)$ can be approximated as

$$P(\mathcal{D} \mid i) \approx \frac{1}{\iota} \sum_{\iota=1}^{\phi} \frac{2 \times P_{i\iota}(\mathcal{D}) \times R_{i\iota}(\mathcal{D})}{P_{i\iota}(\mathcal{D}) + R_{i\iota}(\mathcal{D})} \tag{28}$$

where $P_{i\iota}(\mathcal{D})$ and $R_{i\iota}(\mathcal{D})$ denotes Precision (P) and Recall (R) (Section 3.3.5) obtained by classifier i at fold ι.

According to Equation (27), we take into account the vote of each classifier by exploiting the prediction marginal instead of a 0/1 vote and we tune this probabilistic claim according to the ability of the classifier to fit the training data. This approach allows the uncertainty of each classifier to be taken into account, avoiding over-confident inferences. The BMA paradigm is summarized in Figure 13.

5.3 MODEL SELECTION STRATEGY

A crucial issue of most ensemble methods is referred to the selection of the optimal set of models to be included in the ensemble. This is

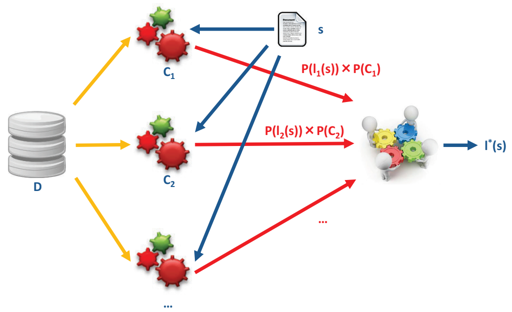


Figure 13: BMA classification

a combinatorial optimization problem over $\sum_{p=1}^{N} \frac{N!}{p!(N-p)!}$ possible solutions where N is the number of classifiers and p represents the dimension of each potential ensemble. Several metrics have been proposed in the literature to evaluate the contribution of classifiers to be included in the ensemble (see (Partalas et al., 2010)). However, these measures are not suitable for a Bayesian Ensemble, because they assume uniform weight distribution of classifiers. For this reason, we propose a heuristic able to compute the discriminative marginal contribution that each classifier provides with respect to a given ensemble. In order to illustrate this strategy, consider a simple case with two classifiers named i and j.

To evaluate the contribution (gain) that the classifier i gives with respect to j, we need to introduce two cases:

1. j incorrectly labels the sentence s, but i correctly tags it. This is the most important contribution of i to the voting mechanism and represents how much i is able to correct j's predictions;

2. Both i and j correctly label s. In this case, i corroborates the hypothesis provided by j to correctly label the sentence.

On the other hand, i could also bias the ensemble prediction in the following cases:

3. j correctly labels sentence s, but i incorrectly tags it. This is the most harmful contribution in a voting mechanism and represents how much i is able to negatively change the (correct) label provided by j.

4. Both i and j incorrectly label s. In this case, i corroborates the hypothesis provided by j leading to a double misclassification of the sentence.

To formally represent the cases above, let compute $P(i = 1 \mid j = 0)$ as the number of instances correctly classified by i over the number of instances incorrectly classified by j (case 1) and $P(i = 1 \mid j = 1)$ the number of instances correctly classified both by i over the number of instances correctly classified by j (case 2). Analogously, let $P(i = 0 \mid$

$j = 1)$ be the number of instances misclassified by i over the number of instances correctly classified by j (case 3) and $P(i = 0 \mid j = 0)$ the number of instances misclassified by i over the number of instances misclassified also by j (case 4).

The contribution r_i^S of each classifier i belonging to a given ensemble $S \subseteq C$ can be estimated as:

$$r_i^S = \frac{\sum\limits_{j \in \{S \setminus i\}} \sum\limits_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum\limits_{j \in \{S \setminus i\}} \sum\limits_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)} \tag{29}$$

where $P(j = q)$ is the prior of classifier j to either correctly or incorrectly predict labels. In particular, $P(j = 1)$ denotes the percentage of correctly classified instances (i. e. accuracy), while $P(j = 0)$ represents the rate of misclassified (i. e. error rate).

Once the contribution of each classifier has been computed, a further issue to be addressed concerns with the search strategy for determining the optimal ensemble composition. The greedy approaches presented in the literature can be distinguished, as reported in Figure 14, according to the search direction: *forward selection* (Fan et al., 2002; Martínez-Muñoz and Suárez, 2006) and *backward elimination* (Banfield et al., 2004; Caruana et al., 2004).

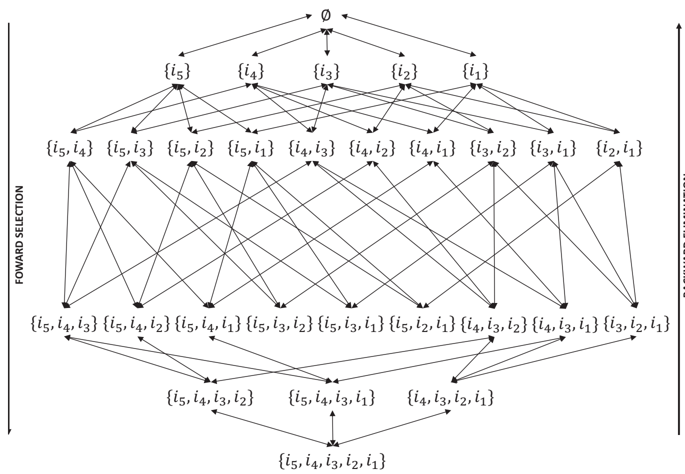


Figure 14: Example of search strategies for an ensemble composed of five models

In *forward selection*, the initial ensemble S is an empty set. The algorithm iteratively adds to S the classifier $i \in \{C \setminus S\}$ that optimizes a given evaluation function. In *backward elimination*, the ensemble S initially contains all the classifiers of the complete set C and iteratively removes the classifier $i \in S$ that optimizes the evaluation function. The advantage of backward elimination is that recognizing irrelevant models is straightforward. Removing a relevant model from a complete set should cause a decline in the evaluation, while adding a relevant model to an incomplete set may have an immediate impact. According to this consideration, the proposed evaluation function r_i^S is included in a greedy strategy based on backward elimination: starting from an initial set $S = C$, the contribution r_i^S is iteratively computed excluding at each step the classifier that achieves the lowest r_i^S. The proposed strategy allows us to reduce the search space from $\sum_{p=1}^{n} \frac{n!}{p!(n-p)!}$ to $n-1$ potential candidates for determining the optimal ensemble, because at each step the classifier with the lowest r_i^S is disregarded until the smallest combination is achieved.

Another issue that concerns greedy selection is the stop condition related to the search process, i.e. how many models should be included in the final ensemble. The most common approach is to perform the search until all models have been removed from the ensemble and select the sub-ensemble with the lowest error (or higher accuracy) on the evaluation set. Alternatively, other approaches select a fixed number of models (or a percentage of the original ensemble).

BMA proposes to perform a backward selection until a local maxima of average classifier contribution is achieved. In particular, the backward elimination will continue until the average classifier contribution (ACC) of a sub-ensemble with respect to the previous parent ensemble will decrease. Indeed, when the average contribution decreases the parent ensemble corresponds to a local maxima and therefore is accepted as optimal ensemble combination. More formally, an ensemble S is accepted as optimal composition if the following condition is satisfied:

$$\frac{ACC(S)}{|S|} \geqslant \frac{ACC(S \setminus x)}{|S-1|} \tag{30}$$

where $ACC(S)$ is estimated as the average r_i^S over the classifiers belonging to the ensemble S. Note that the contribution of each classifier i is computed according to the ensemble S, that is iteratively updated once the worst classifier is removed. This leads to the definition of S characterized by a decreasing size ranging from $|S| = N, N-1, \ldots, 1$.

In order to define the initial ensemble, the baseline classifiers in C have to show some level of dissimilarity. This can be achieved using models that belong to different families (i.e. generative, discriminative and large-margin models). As general remarks, this diversity helps ensembles to better capture different patterns of the natural language. Once this requirement is satisfied, the baseline classifiers to be enclosed in an ensemble can be arbitrary selected.

5.4 TRADITIONAL CLASSIFIER ENSEMBLE

- *Majority Voting (MV)* is the most popular ensemble system. It is characterized by a set of “experts” that classifies the sentence polarity by considering the vote of each classifier as equally important and determines the final polarity by selecting the most popular label prediction (Dietterich, 2002).

Given the set C of independent classifiers and $l_i(s)$ the label assigned to a sentence s by the classifier $i \in C$, the optimal label $l^*(s)$ is assigned as follows:

$$l^*(s) = \begin{cases} \arg\max_k \sum_{i \in C} I(l_i(s) = k) & \text{if } \sum_{i \in C} I(l_i(s) = k) > \\ & \sum_{i \in C} I(l_i(s) = k') \\ & \forall k' \neq k \in K \\ \hat{l}(s) & \text{otherwise} \end{cases} \tag{31}$$

where $I(\cdot)$ is the indicator function, K is the set of labels and $\hat{l}(s)$ is the label assigned to s by the "most expert" classifier, i. e. the classifier that is able to ensure the highest accuracy. In particular, since in this study we consider only two labels $l_i(s)_+$ and $l_i(s)_-$ (positive or negative sentences), the optimal label can be determined according to the following decision rule:

$$l^*(s) = \begin{cases} + & \text{if } \sum_{i \in C} I(l_i(s)_+) > \sum_{i \in C} I(l_i(s)_-) \\ - & \text{if } \sum_{i \in C} I(l_i(s)_+) < \sum_{i \in C} I(l_i(s)_-) \\ \hat{l}(s) & \text{otherwise} \end{cases} \tag{32}$$

Alternative voting systems can be easily derived by exploiting different combination rules based on the posterior probabilities of label prediction provided by the considered models. The most popular combination rules are:

MAXIMUM RULE It selects the maximum a posteriori probability among the classifiers in the ensemble according to:

$$P(l^*(s)) = \max_i P(l_i(s)), \quad i \in C \tag{33}$$

MEAN RULE The decision is determined according to the mean of the a posteriori probabilities given by the classifiers:

$$P(l^*(s) = k) = \frac{1}{|C|} \sum_{i \in C} P(l_i(s) = k) \tag{34}$$

PRODUCT RULE The decision is determined by the product of the posterior probabilities:

$$P(l^*(s) = k) = \prod_{i \in C} P(l_i(s) = k) \tag{35}$$

- *Bagging* (Quinlan, 1996; Bauer and Kohavi, 1999) is another very popular ensemble technique also approached for polarity classification. The main goal of Bagging is to aggregate the multiple hypotheses generated by the same learning algorithm on different distributions of training data.

  Bagging assumes a given training set $\mathcal{D}$ and a learning system which trains a base classifier for each training set (i.e. bags) $b = 1, 2, \ldots, B$ sampled with replacement from $\mathcal{D}$. This implies that some instances may not appear in a given bag, while others appear more than once. The learning system trains a base classifier to infer the label $l_b(s)$ for each sentence of the testing set.

  The predictions $l_b(s)$ are then aggregated over all the bags according to a *MV* decision rule:

$$l^*(s) = \begin{cases} \arg\max_k \sum_{b=1}^{B} I(l_b(s) = k) & \text{if } \sum_{b=1}^{B} I(l_b(s) = k) > \\ & \sum_{b=1}^{B} I(l_b(s) = k'), \\ & \forall k' \neq k \in K \\ \widehat{l}(s) & \text{otherwise} \end{cases} \tag{36}$$

where $\widehat{l}(s)$ is the label assigned to s by the "most expert" classifier, i. e. the classifier that is able to ensure the highest accuracy.

Other combination rules based on the posterior probabilities $P(l_b(s) = k)$ can be used to derive the final optimal label of a given sentence. Analogously to simple voting systems, the most popular decision rules that have been investigated into the experimental phase are: *maximum*, *mean* and *product* rule.

MAXIMUM RULE It selects the maximal a posteriori probability of the classified sentence s among bags:

$$P(l^*(s)) = \max_b P(l_b(s)), \quad b = 1, 2, \ldots, B \tag{37}$$

MEAN RULE The decision is determined according to the mean of the posterior probabilities among bags:

$$P(l^*(s) = k) = \frac{1}{B} \sum_{b=1}^{B} P(l_b(s) = k) \tag{38}$$

PRODUCT RULE The decision is determined by the product of the a posteriori probabilities

$$P(l^*(s) = k) = \prod_{b=1}^{B} P(l_b(s) = k) \tag{39}$$

Considering that Bagging depends on a random sampling of training instances of the original dataset, 10 execution runs have been performed. Each run has enclosed 9 bags for inducing each baseline classifier.

5.5 EXPERIMENTAL INVESTIGATION

In order to evaluate the contribution of the proposed Bayesian Ensemble Learning, a comparison with Majority Voting, Bagging and traditional baseline classifiers of state of the art is proposed. The studied classifiers are: *Naïve Bayes (NB)* (McCallum and Nigam, 1998), *Support Vector Machines (SVM)* (Cortes and Vapnik, 1995), *Maximum Entropy (ME)* (McCallum et al., 2006), *Conditional Random Fields (CRF)* (Sutton and McCallum, 2012) and *DIC* (Hu and Liu, 2004).

5.5.1 *Dataset and Evaluation Criteria*

Several benchmarks have been considered for evaluation. The first evaluation is based on *Review* datasets:

- *Sentence polarity dataset v1.0*, in the following denoted as **MovieData**[1]. This dataset, presented in (Pang and Lee, 2005), is composed of 10662 snippets of movie reviews extracted from Rotten Tomatoes[2]. The main characteristics of this dataset, which comprises only positive and negative sentences, are related to the informal language adopted (slang and short forms) and to the presence of noisy polarity labeling.

- *Finegrained Sentiment Dataset, Release 1*, in the following denoted as **ProductData**[3]. This dataset, originally presented in (Täckström and McDonald, 2011), relates to product reviews about books, dvds, electronics, music and video games taken from `Amazon.com`. Although the original dataset is composed of 5 polarities, a reduction of instances has been performed to deal only with positive and negative opinions. The resulting dataset is unbalanced, composed of 1320 ($\simeq 58.84\%$) negative and 923 ($\simeq 41.16\%$) positive reviews.

- *Multi-Domain Sentiment Dataset*, in the following denoted as **ProductDataMD**[4]. This dataset, originally introduced in (Blitzer et al., 2007), contains product reviews taken from `Amazon.com` about many product types (domains). Reviews contain star ratings (1 to 5 stars) that have been converted into nominal labels (NEG for ratings lower than 3, NEU for ratings equal to 3

1 http://www.cs.cornell.edu/people/pabo/movie-review-data/
2 http://www.rottentomatoes.com/
3 http://www.sics.se/people/oscar/datasets/
4 http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

and POS for ratings greater than 3). In this study, reviews from category 'Music' and 'Books' are studied separately. Product-DataMD is balanced, composed of 2000 reviews for each of the two categories.

The second type of evaluation is based on *Social* datasets collected from Twitter, i. e. *Gold Standard* benchmark (Chen et al., 2012). The dataset has been distinguished in **Person** and **Movie** according to their main topic. Each gold standard contains 1500 manually labeled Twitter data. Also in this case, although the original dataset is composed of 3 different polarities (POS, NEG and NEU), a reduction of instances has been performed in order to deal only with positive and negative opinions. The resulting datasets are therefore unbalanced:

- **Gold Standard Person**: 105 ($\simeq$ 26.44%) negative and 292 ($\simeq$ 73.56%) positive opinions.
- **Gold Standard Movie**: 96 ($\simeq$ 18.6%) negative and 420 ($\simeq$ 81.4%) positive orientations.

As far is concerned the criteria to evaluate the performance achieved by the investigated methodologies, a 10-folds cross validation has been adopted. As evaluation we adopted the performance measure presented in Section 3.3.5: Precision, Recall, F1-measure and Accuracy. Additionally, we have computed the *Relative Improvement (RI)* to test which Bagging combination rules improves or degrades the performance of the baseline classifiers:

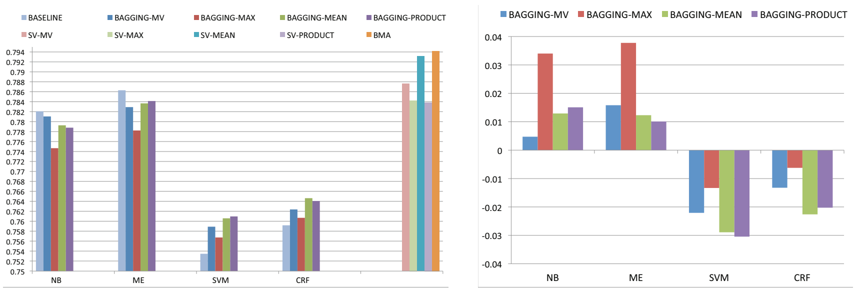
$$\mathrm{RI} = \frac{\mathsf{ERR}_{\mathrm{Bagging}} - \mathsf{ERR}_{\mathrm{base}}}{\mathsf{ERR}_{\mathrm{base}}} \tag{40}$$

where $\mathsf{ERR}_{\mathrm{base}}$ and $\mathsf{ERR}_{\mathrm{Bagging}}$ are the validation error rate for the baseline and the Bagging classifiers, respectively. Considering that RI ranges in the interval [-1,+∞), a negative value indicates that Bagging classifier has a decreasing error rate with respect to baseline classifier.

5.5.2 *Results*

MOVIEDATA Bagging performance on MovieData, compared with the other approaches, are depicted in Figure 15(a) (Simple Voting, denoted with SV-, and BMA bars are the rightmost blocks and are related to the best ensemble composition).

A first interesting observation relates to the comparison between Simple Voting and Bagging: Simple Voting produces higher prediction accuracy than all the Bagging combination rules. Moreover, as highlighted in Figure 15(b), Bagging is highly sensitive to the combination rule with respect to the classifier to be induced. While Bagging with SVM and CRF achieves high RI through the MEAN and PRODUCT combination rule, for NB and ME the base classifiers perform



(a) Bagging vs. Simple Voting vs BMA (b) RI of Bagging combination rules

Figure 15: Bagging Performance on MovieData

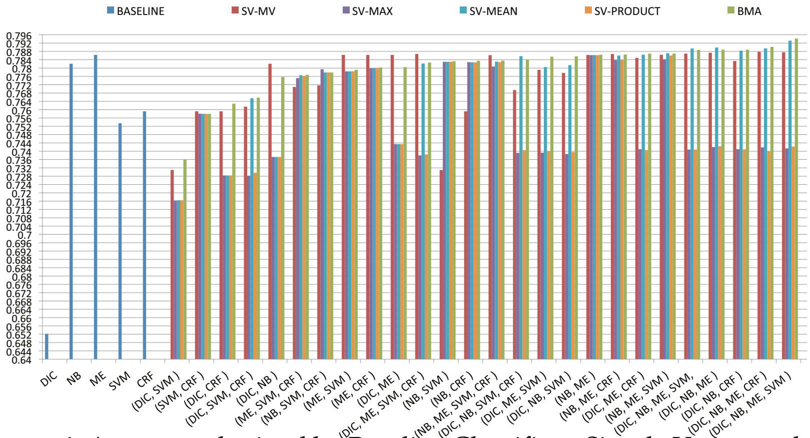


Figure 16: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on MovieData

better than Bagging. A complete comparison of Bagging vs Baseline Classifiers is reported in Appendix A.

In order to compare Baseline Classifiers, Simple Voting and BMA, a summary of accuracy improvements is depicted in Figure 16 (Bagging is omitted because always outperformed by the other ensembles). We can note that DIC obtains low performance on this dataset due to the dictionary composition: the opinion words belonging to the dictionary are concerned with the product reviews and do not fit the movie review of this dataset, leading to poor performance. Although DIC is the worst classifier, most of the promising ensembles contain the dictionary-based classifier. This highlights that the best ensemble is not necessarily composed of the classifiers that individually obtain the highest accuracy: the proposed approach is able to take into account the contribution of DIC classifier by considering it's reliability and prediction, leading to robust and accurate polarity prediction. In particular, the optimal ensemble provided by BMA that includes DIC, NB, ME and SVM, outperforms SV- ensembles achieving 79.41% of accuracy against 78.76% of MV, 74.13% by MAX, 79.31% by MEAN and 74.22% by PRODUCT.

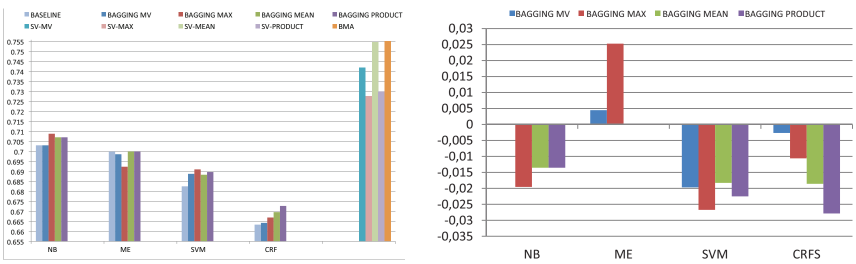
Concerning the proposed model selection strategy, the contribution of each classifier can be computed as shown in Table 4, where the steps of the proposed greedy model selection, classifiers contributions r_i^S, Average Classifier Contribution (denoted as **ACC**) and Accuracy are reported.

Step	DIC	NB	ME	SVM	CRF	ACC	Accuracy
1	1.6618	1.9402	1.9294	1.6740	**1.6486**	1.7709	0.7887
2	**1.6662**	1.9747	2.0042	1.7486	-	**1.8485**	**0.7941**
3	-	1.5868	1.6073	**1.4891**	-	1.5611	0.7869
4	-	**1.1566**	1.2102	-	-	1.1835	0.7863

Table 4: Model Selection on MovieData. Bold-faced numbers denote the contribution r_i^S of the worst classifiers that will be consequently removed from the ensemble.

Starting from the initial set S={DIC, NB, ME, SVM, CRF}, the contribution r_i^S is computed for each classifier. The model with the lowest contribution at the first step is CRF. Then, r_i^S is re-computed on the ensemble $\{S \setminus \text{CRF}\}$, highlighting DIC as the classifier with the lowest contribution. At step 3 and 4, the classifiers to be removed are SVM and NB respectively. It can be easy to note that the model selection strategy has radically reduced the search space and, thanks to the convergence criteria, the optimal BMA ensemble has been ensured at step 2.

PRODUCTDATA Figure 17 shows Simple Voting, Bagging and BMA accuracy achieved on ProductData assuming the Boolean weighing schema. While Bagging (MAX) with NB and SVM obtains higher performance than the corresponding baselines, for ME is the contrary.



(a) Bagging vs. Simple Voting vs BMA (b) RI of Bagging combination rules

Figure 17: Bagging Performance on ProductData

The best Bagging for ProductData is evidently different from the one for MovieData: while for MovieData the optimal Bagging was based on ME and PRODUCT combination rule, for ProductData is given by NB and MAX combination rule. This confirms that Bagging represents a weak ensemble technique. The outperforming ensemble is obtained by BMA leading to small set of expert. Our approach,

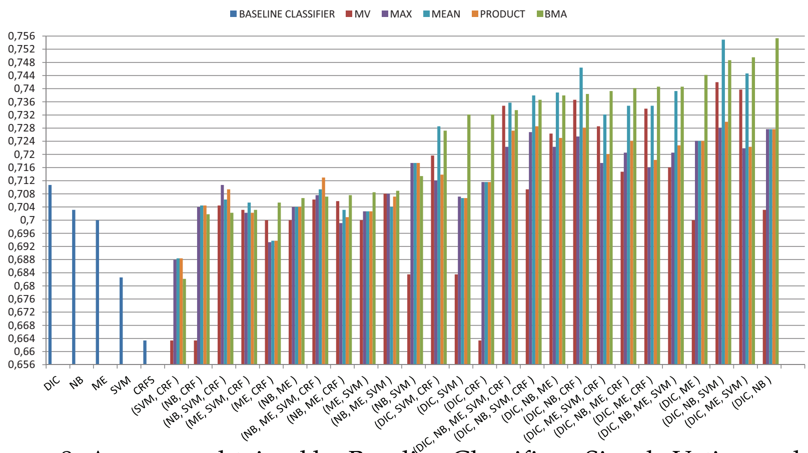


Figure 18: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on ProductData

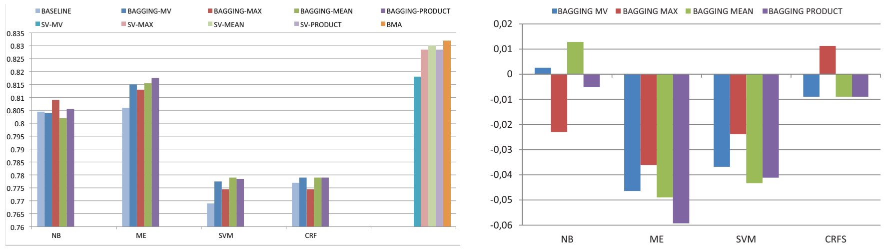
based on DIC and NB, is able to achieve 75.53% of accuracy against 70.31% by MV, 72.76% by MAX, 72.76% by MEAN and 72.76% by PRODUCT obtained by Bagging compositions based on the same experts (all ensemble performance are depicted in Figure 18).

Also in this case, optimal ensemble can be easily identified through the model selection strategy. Starting from the initial ensemble S={DIC, NB, ME, SVM, CRF}, the classifiers are sorted with respect to their contribution by computing Equation (29). As shown in Table 5, the classifier with the lowest contribution at the first step is CRF. Then, r_i^S is re-computed on the ensemble $\{S \setminus \text{CRF}\}$, highlighting ME as the classifier with the lowest contribution. At step 3 and 4, the worst classifiers to be removed from the ensemble are SVM and NB respectively. Once more, the model selection strategy ensures the best BMA ensemble: the greedy search reveals the optimal ensemble at the last step when a local optimum of ACC is achieved.

Step	DIC	NB	ME	SVM	CRF	ACC	Accuracy
1	2.1192	1.7118	1.6568	1.6244	**1.4389**	1.7102	0.7334
2	2.0972	1.6928	**1.6392**	1.6439	-	1.7603	0.7486
3	2.0903	1.9225	-	**1.6894**	-	1.9007	0.7495
4	2.1169	**2.0294**	-	-	-	**2.0731**	**0.7553**

Table 5: Model Selection on ProductData. Bold-faced numbers denote the contribution r_i^S of the worst classifiers that will be consequently removed from the ensemble.

PRODUCTDATAMD (BOOKS) Figure 19 shows Bagging, Simple Voting and BMA performance achieved on ProductDataMD (books). In Figure 19(a) we can easily highlight that, also for this dataset, Simple Voting achieves better accuracy than any Bagging combination rule. As discussed for other datasets, Bagging does not guarantee robust performance: ME and SVM achieve high performance through the PRODUCT and MEAN combination rules, while NB and CRF provide poor results through the MV, MEAN and MAX combination rules.



(a) Bagging vs. Simple Voting vs BMA (b) RI of Bagging combination rules

Figure 19: Comparison between Bagging and ensemble of different classifiers on ProductData (books)

In Figure 20, it can be easy to note that the outperforming ensemble is obtained by BMA with DIC, NB and SVM (83.20% of accuracy). Other approaches obtain accuracy ranging from 80.90% of MV, 82.15% of MAX, 83% of MEAN and 82.45% of PRODUCT following a Bagging paradigm.

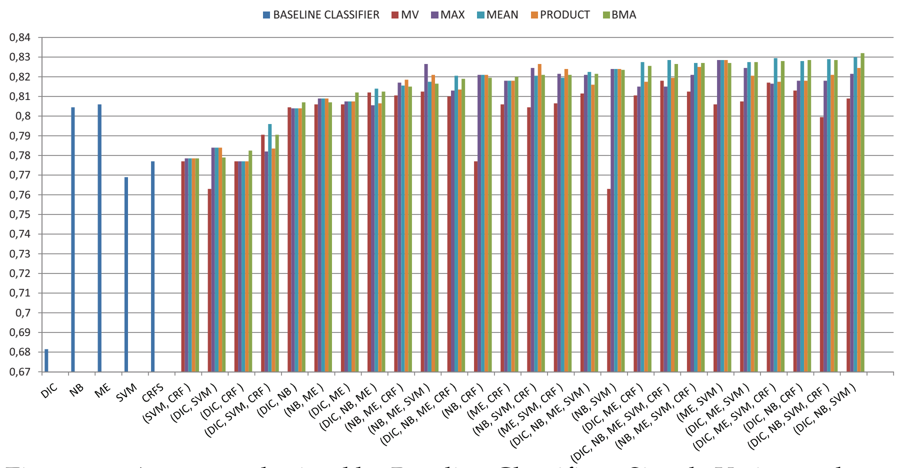
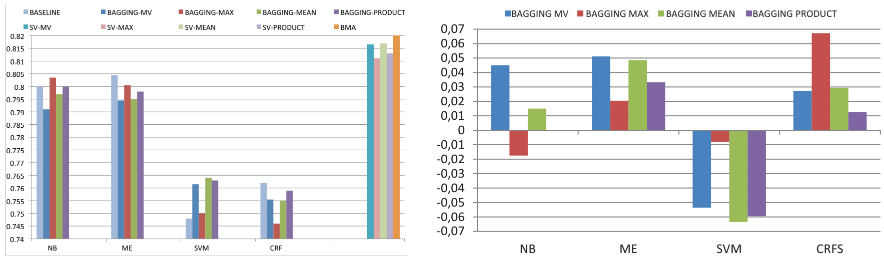


Figure 20: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on ProductDataMD (Books)

The model selection computation (omitted to avoid verbosity) ensures, also for ProductDataMD (Books) dataset, the composition of the best BMA ensemble during the backward elimination process.

PRODUCTDATAMD (MUSIC) Figure 21 shows Bagging versus the other approaches on ProductDataMD(Music). In this case, Bagging-MEAN achieves high performance through SVM, while for NB, ME and CRF the base classifiers perform better than Bagging. Concerning the combination rules enclosed in Bagging, the results about error relative improvement show low-quality performance. Simple Voting is confirmed to be a promising approach to be compared with the proposed BMA.



(a) Bagging vs. Simple Voting vs BMA (b) RI of Bagging combination rules

Figure 21: Bagging performance on ProductDataMD (Music)

In order to compare Baseline Classifiers, Simple Voting and BMA, a summary of accuracy improvements is depicted in Figure 22 (also in this case Bagging is omitted because always outperformed by the other ensembles). Concerning ProductDataMD (Music) the accuracy measure reveals that the outperforming ensemble, obtained by the BMA paradigm, comprises all the baseline classifiers (NB, ME, SVM, CRF and DIC). The proposed approach is able to achieve 82% of accuracy against the Simple Voting systems (81.45% by MV, 80.65% by MAX, 81.7% by MEAN and 80.9% by PRODUCT).

Also in this case, the model selection strategy ensures the composition of the best BMA ensemble during the greedy search, radically reducing the space search.

GOLD STANDARD PERSON Figure 23 shows Bagging versus the other approaches on Gold Standard Person. In this case, Bagging achieves high performance through NB considering all the combination rules, while for SVM only the MEAN and PRODUCT rules achieve highest results through Bagging. For ME and CRF, the base classifiers perform better than Bagging.

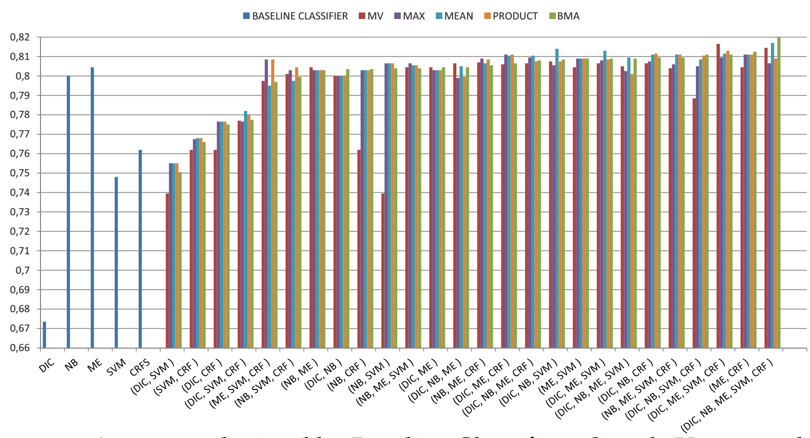
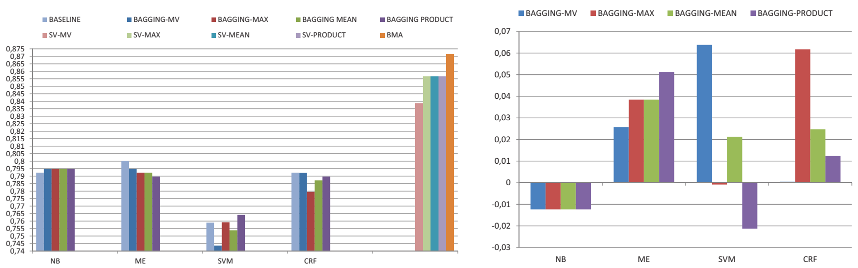


Figure 22: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on ProductDataMD (Music)



(a) Bagging vs. Simple Voting vs BMA

(b) RI of Bagging combination rules

Figure 23: Bagging performance on Gold Standard Person

Concerning the combination rules enclosed in Bagging, the results about error relative improvement show low-quality performance. Simple Voting is confirmed to be a promising approach to be compared with the proposed BMA.

In order to compare Baseline Classifiers, Simple Voting and BMA, a summary of accuracy improvements is depicted in Figure 24 (also in this case Bagging is omitted because always outperformed by the other ensembles). Concerning Gold Standard Person, the accuracy measure reveals that the outperforming ensemble, obtained by the BMA paradigm, comprises only DIC and NB. The proposed approach is able to achieve 87.13% of accuracy against the Simple Voting systems (83.84% by MV, 85.64% by MAX, MEAN and PRODUCT).

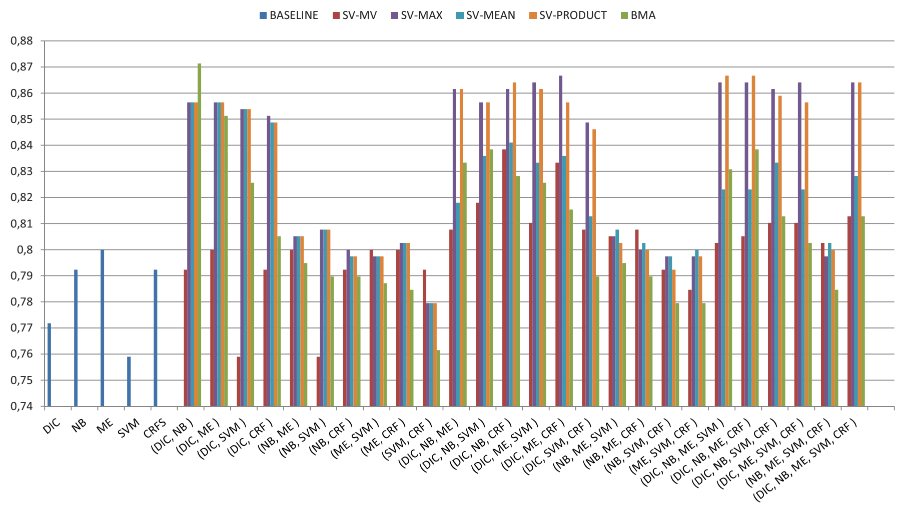
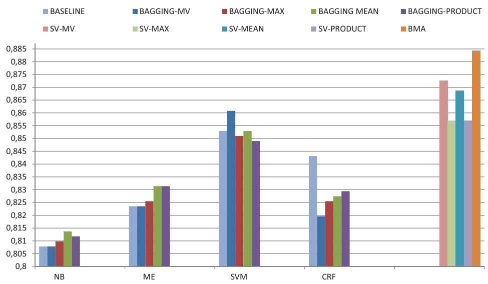
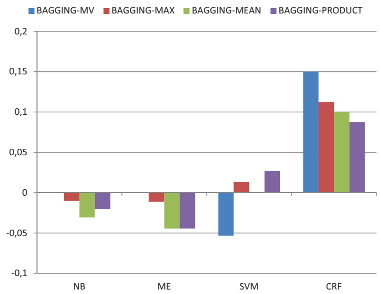


Figure 24: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on Gold Standard Person

GOLD STANDARD MOVIE Regarding Gold Standard Movie, analogous results have been obtained for the optimal weighting schema. The comparison between Bagging and the other approaches on Gold Standard Movie is shown in Figure 25.



(a) Bagging vs. Simple Voting vs BMA



(b) RI of Bagging combination rules

Figure 25: Bagging performance on Gold Standard Movie

In this case, Bagging achieves high performance through NB and ME with all the considered combination rules, while for CRF the base classifiers perform better than Bagging. Using SVM, Bagging-MV achieves the highest performance. Concerning the combination rules enclosed in Bagging, the results about error relative improvement show high-quality performance, except for CRF.

A summary of accuracy improvements is depicted in Figure 26, where the outperforming ensemble obtained by BMA comprises DIC, SVM and CRF.

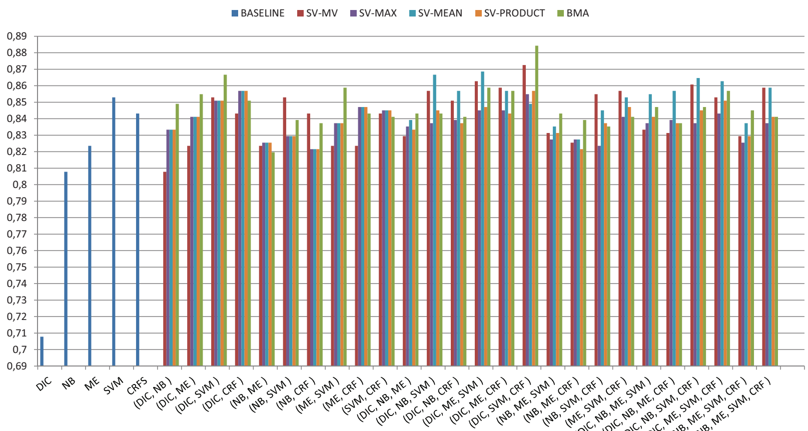


Figure 26: Accuracy obtained by Baseline Classifiers, Simple Voting and BMA on Gold Standard Movie

The proposed approach is able to achieve 88.43% of accuracy against the Simple Voting systems (87.25% by MV, 85.68% by MAX, 86.86% by MEAN and 85.68% by PRODUCT). The model selection strategy, applied to both Gold Standard Movie and Person, ensures again the composition of the optimal BMA ensemble through the greedy search.

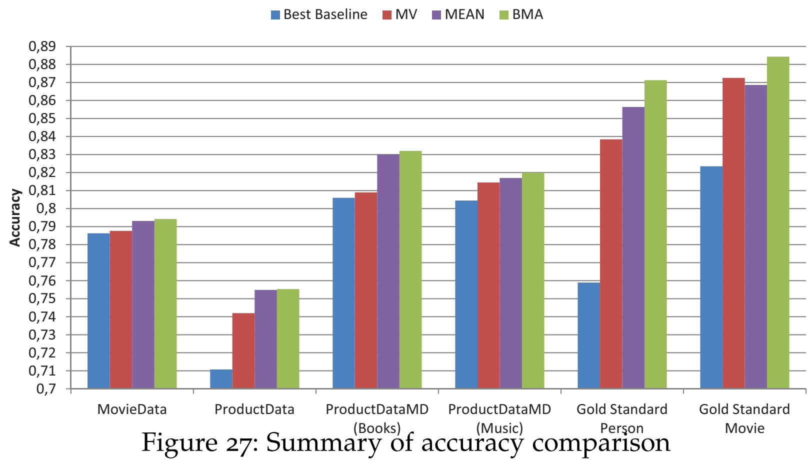


Figure 27: Summary of accuracy comparison

5.6 Computational complexity analysis

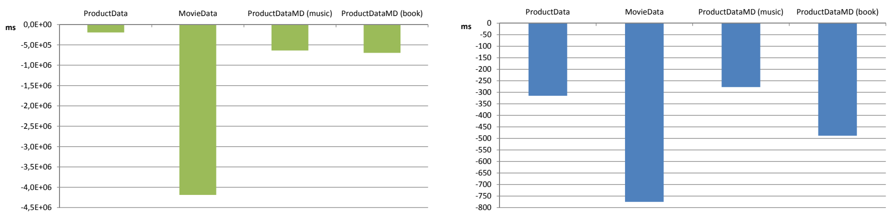
In order to demonstrate the efficiency of the proposed approach compared to the other ensemble techniques, a computational complexity analysis has been performed. As far is concerned with MV approach, it is well known that it belongs to the NP-Hard complexity class due to the exhaustive search (combinatoric problem) to find the optimal ensemble composition. Bagging, as bootstrapping technique, is a computationally intensive procedures mainly affected by the training phase. The resampling step, together with the learning of a given classifier for a given number of bags, lead to a time expensive approach. On the contrary, BMA results in a more efficient paradigm characterized by reduced time costs. To better grasp the computational complexity of the considered ensemble techniques, we can distinguish between three phases: training, search of the optimal ensemble and inference. While the computational complexity of MV and BMA mainly depends on the number N of classifiers to be enclosed in the ensemble, Bagging takes also into account the number of replacements R and the number of bags B. Assuming that learning and inference on a given classifier takes $\mathcal{O}(1)$, a comparison of the time complexity is reported in Table 6.

	Training	**Search**	**Inference**
MV	$\mathcal{O}(N)$	$\mathcal{O}(2^N)$	$\mathcal{O}(N)$
Bagging	$\mathcal{O}(B(N+R))$	$\mathcal{O}(N)$	$\mathcal{O}(B)$
BMA	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$

Table 6: Computational Complexity of Ensemble Learning Techniques

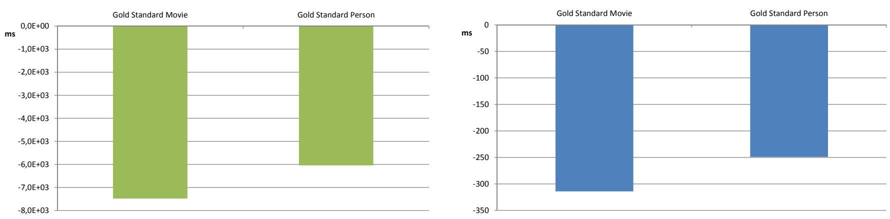
Concerning the training phase, although Bagging is characterized by a linear time complexity as well as the two other approaches, it results to be the most computationally intensive technique in practice. Indeed, while BMA and MV can be solved in $\mathcal{O}(N)$, Bagging results in a higher computational complexity equal to $\mathcal{O}(B(N+R))$ due to B and R that are by definition greater than one. Regarding the search of the optimal ensemble composition, MV is the most time consuming approach characterized by an exponential time complexity of $\mathcal{O}(2^N)$ compared to the linear BMA and Bagging. Indeed, while MV must search the optimal ensemble over an hypothesis space of $\sum_{p=1}^{N} \frac{N!}{p!(N-p)!} = 2^{N-1}$, Bagging has to search over N possible candidates (all the weak classifiers are candidate to be bootstrapped) and BMA over $N-1$ candidate sub-ensembles. In the inference phase, all the ensemble learning approaches results to be linear in time complexity. However, MV and BMA are more efficient when the number of bags B is greater than the number of classifiers N enclosed in the ensemble, i. e. $B > N$. From a theoretical point of view, for an increasing number of classifiers to be enclosed in an ensemble, BMA results to be the most efficient approach.

In order to better grasp the efficiency of the proposed approach, a further analysis has been performed. In Figure 28 (*Review* benchmarks) and Figure 29 (*Social* Datasets) the time reductions of BMA in respect of the other approaches are reported considering the three phases, i. e. training, search of the optimal ensemble and inference[5].



(a) BMA vs Bagging (b) BMA vs MV

Figure 28: Efficiency comparison (in terms of milliseconds) on *Review* benchmarks



(a) BMA vs Bagging (b) BMA vs MV

Figure 29: Efficiency comparison (in terms of milliseconds) on *Social* benchmarks

If we focus on Figure 28, we can easily observe that BMA ensures a valuable time reduction for all the considered *Review* datasets. When

5 Time performance have been measured on a Desktop PC with Windows 7 64-bit Operating System, Pentium Quad Core i7 3.10GHz Processor and 8GB RAM.

dealing with large datasets, as for example MovieData, the gain becomes more evident: while for the entire process BMA takes 8.41 min, Bagging performs in 78.22 min and MV in 8.43 min. On *Social* datasets (Figure 29), the time reductions provided by BMA is lower than *Review* benchmarks. This is due both to a smaller set of instances and to the short nature of text (posts on Twitter are limited to 140 characters). However, also in this case BMA guarantees a time reduction: on Gold Standard Person, BMA takes 1151 ms against 1400 of MV and 7194 of Bagging, while on Gold Standard Movie, it takes 1353 ms against 1667 of MV and 8835 of Bagging.

In conclusion, we argue that BMA, together with the proposed model selection strategy, ensures a significant performance improvement with regard to the studied baseline classifiers and the Simple Voting. BMA outperforms not only any Bagging composition obtained with the considered combination rules, but also the best baseline classifiers and Simple Voting (i. e. SV-MEAN). Moreover, BMA works better both on well-formed documents (e. g., reviews) and social network messages (e. g., tweets). Finally, experimental results show that the proposed solution is particularly effective and efficient, thanks to its ability to define a strategic combination of different classifiers through an accurate and computationally efficient heuristic, that is particularly challenging when dealing with online and real-time big data. It is important to remark that the proposed approach radically reduces the search space for composing the optimal ensemble, ensuring outperforming performance than other standard approaches.

BMA is particularly effective and efficient

A further contribution based on BMA relates to the classification of subjective/objective (Section 3.1.3.2) and ironic/non ironic messages (Section 3.1.3.6). In the literature all these tasks related to Sentiment Analysis are usually tackled independently on each other. In a real context all these issues should be addressed by a single model able to distinguish at first if a message is either subjective or objective, to subsequently address polarity and irony detection and deal with the potential relationships that could exists between them. For this reason, a Hierarchical Bayesian Model Averaging (H-BMA) framework able to jointly address subjectivity, polarity and irony detection is proposed (Fersini et al., 2014c).

6 EXPLOITING NETWORK INFORMATION FOR POLARITY CLASSIFICATION: DOCUMENT AND USER-LEVEL

"Relationships are all there is.
Everything in the universe only exists
because it is in relationship to
everything else. Nothing exists in
isolation. We have to stop pretending we
are individuals that can go it alone."
— M. Wheatley,
Writer

In this chapter, a particular type of Probabilistic Relational Model (Section 4.5) whose goal is to estimate the user-level polarity is proposed. It is inspired to the Markov Random Field, a set of random variables having a Markov property described by an undirected graph.

When dealing with sentiment classification on social networks, a fully supervised learning paradigm is usually required, where the sentiment orientation of users must be known a priori to derive suitable predictive models. However, this does not reflect the real setting of social networks, where the polarity on a given topic is explicitly available only for some users (black nodes) while for others could be derived from their posts and relations with other users (white nodes). As black nodes we considered those users whose bio (description in Twitter) or name clearly state a positive or negative opinion about a given topic. For instance, regarding the topic 'Obama', a positive user's bio could report "*I love football, TV series and Obama!*" and/or the name could be "*ObamaSupporter*". In this context, a semi-supervised learning paradigm better represents the real setting.

Moreover, most of the works in SA (Go et al., 2009; Barbosa and Feng, 2010) are merely based on textual information expressed in microblogs. However, they do not consider that microblogs are actually networked environments. As mentioned in Chapter 4, early studies for overcoming this limitation exploit the principle of homophily (McPherson et al., 2001) for dealing with user connections. This principle, i.e. the idea that similarity and connection between users tend to co-occur, could suggests that users connected by a personal relationships may tend to hold similar opinions. According to this social principle, friendship relations have been considered in few recent studies. In (Tan et al., 2011; Hu et al., 2013), authors explore user-level sentiment analysis by modelling tweet contents and friendship relationships among users. However, in (Tan et al., 2011) tweet labels are not estimated but given a priori according to the real user label, and the mention network (a user mentions another one using the Twitter @-convention) is also considered following the motivation that users will mention those who they agree with. Although social relations play a fundamental role in SA on microblogs, we argue that considering friendship connections is a weak assumption for modelling homophily: two friends might not share the same opinion about a given topic.

To these purposes, we present a semi-supervised sentiment learning approach (Pozzi et al., 2013c) able to deal both with text and Approval Network (Chapter 4): given a small proportion of users already labeled in terms of polarity, it predicts the sentiments of the remaining unlabeled users by combining textual information and Approval Network directly in the probabilistic model.

6.1 THE CLASSIFICATION MODEL

Given a H-DAG where approval edges are weighted through Equation (23), two vectors need to be introduced to tackle the sentiment classification problem at user-level: a vector of labels $\mathbf{L}^V = \{l(v_i) \in \{+,-\} | v_i \in V\}$ that defines each user as either "positive" (+) or "negative" (-) and an analogous vector of labels $\mathbf{L}^M = \{l_{v_i}(m_t) \in \{+,-\} | v_i \in V, m_t \in M\}$ that represents the polarity label of each message m_t written by user v_i. Since the sentiment $l(v_i)$ of the user v_i is influenced by the sentiment labels $l_{v_i}(m_t)$ of his messages and the sentiment labels of the directly connected neighbours $N(v_i)$, the approval model is intended to obey to the Markov assumption.

The user-message and user-user (approval) relations are combined in the **Semi-supervised Sentiment Learning by Approval Network (S²-LAN)** as follows:

$$
\begin{aligned}
&\log P(\mathbf{L}^V \mid M, \phi) = \\
&\Bigg(\sum_{v_i \in V} \Bigg[\sum_{m_t \in M} \sum_{\alpha} \sum_{\beta} \mu_{\alpha,\beta} f_{\alpha,\beta}(l(v_i), l_{v_i}(m_t)) \\
&+ \sum_{\substack{v_j \in N(v_i), \\ (v_i, v_j) \in \phi}} \sum_{\alpha} \sum_{\beta} \lambda_{\alpha,\beta} g_{\alpha,\beta}(l(v_i), l(v_j)) \Bigg] \Bigg) - \log Z
\end{aligned}
\tag{41}
$$

where $f_{\alpha,\beta}(\cdot,\cdot)$ and $g_{\alpha,\beta}(\cdot,\cdot)$ are feature functions that evaluate user-message and user-user relations respectively ($\alpha, \beta \in \{+,-\}$), and the weights $\mu_{\alpha,\beta}$ (a user with label α who posts a message with label β), $\lambda_{\alpha,\beta}$ (a user with label α connected to a user with label β) need to be estimated. Z is the normalization factor that enables a coherent probability distribution of $P(\mathbf{L}^V)$. Regarding the estimation of μ, λ and the assignment of user sentiment labels which maximizes $\log P(\mathbf{L}^V)$ refer to (Pozzi et al., 2013c), where a modified version of SampleRank Algorithm (Wick et al., 2009) is presented.

USER-MESSAGE FEATURE FUNCTION A user-message feature function evaluates whether the message polarity agrees (or disagrees) with respect to the user sentiment. Formally, $f_{\alpha,\beta}(l(v_i), l_{v_i}(m_t))$ is defined as:

$$f_{\alpha,\beta}(l(v_i), l_{v_i}(m_t)) =$$
$$\begin{cases} \frac{\rho_{T-black}}{|M_{v_i}|} & l(v_i) = \alpha, l_{v_i}(m_t) = \beta, v_i \text{ black} \\ \frac{\rho_{T-white}}{|M_{v_i}|} & l(v_i) = \alpha, l_{v_i}(m_t) = \beta, v_i \text{ white} \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

where "v_i black" means that user v_i is initially labeled (i. e. its polarity label is known a priori), and "v_i white" means that user v_i is unlabeled (i. e. its polarity label is unknown a priori). The parameters $\rho_{T-black}$ and $\rho_{T-white}$ represent the different level of confidence in black and white users[1], and $M_{v_i} \subset M$ denotes the set of messages written by user v_i. Every $m_t \in M$ is assumed to have a polarity label. Bayesian Model Averaging (Section 5.2) is used to automatically label messages.

USER-USER FEATURE FUNCTION A user-user feature function evaluates whether the polarity of a given user agrees (or disagrees) with its neighbour's sentiment. Given a H-DAG (ϕ) defined in Def. 4.4 (Section 4.4), $g_{\alpha,\beta}(l(v_i), l(v_j))$ is formally defined as follows:

$$g_{\alpha,\beta}(l(v_i), l(v_j)) = \begin{cases} \frac{\rho_{neigh} \cdot c_{i,j}}{\sum_{v_k \in N(v_i)} c_{i,k}} & l(v_i) = \alpha, l(v_j) = \beta \\ 0 & \text{otherwise} \end{cases} \tag{43}$$

where ρ_{neigh} represents the level of confidence in relationships among users[1] and $c_{i,j}$ denotes the normalized weights of approval edges in ϕ.

6.2 EXPERIMENTAL INVESTIGATION

In this section, we present a case study to validate the semi-supervised model S^2-LAN presented above. In particular, the experimental investigation is focused on connections derived from Twitter and presents a comparison between the ability of inferring user-level polarity classification and traditional approaches based only on textual features.

1 Note that $\rho_{T-black}$, $\rho_{T-white}$ and ρ_{neigh} are empirically estimated (see Section 6.2.1).

6.2.1 *Dataset and Settings*

The network for evaluating the investigated model comprises 62 users[2] posting about Obama, whose tweets have been monitored during the period 8-10 May 2013. The resulting tweets have been manually labeled as positive or negative by three annotators as well as the related users. Concerning the message polarity classification, BMA model has been trained by using positive and negative tweets of the *Obama-McCain Debate (OMD)*[3] dataset. S^2-LAN has been compared with traditional approaches based on text, i. e. Dictionary-based classifier (DIC) (Hu and Liu, 2004), Naïve Bayes (NB) (McCallum and Nigam, 1998), Maximum Entropy (ME) (McCallum et al., 2006), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Conditional Random Fields (CRF) (Sutton and McCallum, 2012) and Bayesian Model Averaging (BMA) (Pozzi et al., 2013c). The classical state-of-the-art measures for classification described in Section 3.3.5 have been employed, i. e. Precision (P), Recall (R), F1-measure (F1) and Accuracy (Acc).

6.2.2 *Results*

The first evaluation relates to the ability of classifiers to detect the polarity of tweets. By comparing the state-of-the-art approaches, it emerges that BMA achieves 60.37% of Accuracy compared with 58.49% obtained by the best single classifier (CRF). According to these results, BMA has been selected as text-based baseline for the comparison with S^2-LAN. In particular, the polarity of users has been derived by aggregating their tweets through a majority voting mechanism. For instance, if BMA detects three positive and two negative tweets for a given user, the final user label will be 'positive'. When inferring polarities on unlabeled users (white nodes), this heuristic achieves a 66.66% of Accuracy. S^2-LAN based on Approval Network is able to achieve 93.8% of Accuracy, significantly outperforming the BMA-based approach.

Table 7 summarizes the performance achieved by S^2-LAN with respect to the BMA approach. Note that approval relations enclosed in S^2-LAN ensure a global improvement of 27% with respect to the text-only method. Since BMA does not take into account any kind of relationship, the correct prediction of a user does not have any effect on adjoining users.

Considering the network, the prediction of each user has impact on all the other nodes by a "propagation" effect, smoothing each predicted label according to adjoining nodes. This investigation finally confirms that the inclusion of relationships in predictive models, as

2 Starting from an initial set of 2500 users, only users with positive and negative tweets become part of the final set

3 https://bitbucket.org/speriosu/updown/src/5de483437466/data/

	P_-	R_-	$F1_-$	P_+	R_+	$F1_+$	Acc
BMA	0.655	0.826	0.731	0.692	0.474	0.563	0.666
S^2-LAN	0.933	0.963	0.945	0.958	0.907	0.925	0.938
Gain	+0.278	+0.137	+0.214	+0.266	+0.433	+0.357	+0.27

Table 7: Performance achieved considering only text (BMA) and relations + text (S^2-LAN)

suggested in other studies (Fersini et al., 2010a; Sharara et al., 2011; Fersini and Messina, 2013), leads to improve recognition performance when dealing with non-propositional environments.

7 ASPECT-SENTIMENT MODELING USING NETWORK

"When we care, we share."
— Jonah Berger,
Contagious: Why Things Catch On

This chapter is focused on simultaneously extracting aspects and classifying sentiments from textual messages. To this issue, an unsupervised type of Probabilistic Relational Model (Section 4.5) called Networked Aspect-Sentiment model (NAS) is proposed.

7.1 BACKGROUND

Most of the works in SA (Wang and Manning, 2012; Maas et al., 2011; Go et al., 2009; Barbosa and Feng, 2010) are topic-dependent, i. e. they identify the sentiments of documents given a particular topic filtered a priori. On Social Media it is usually extracted through social tools such as "hashtags" (words or unspaced phrases prefixed with the symbol '#'). However, this is not always sufficient, because hashtags usually identify the overall topic of a text message and not its sub-topics (or aspects). For example, the above works would assign the topic 'iOS7' to the tweet *"#iOS7 is very good, but they still need to work on battery life and security issues"*, despite the sentiment can be distinguished between the aspects 'battery life' and 'security'. As already described in Section 3.4, *Topic Sentiment Mixture model (TSM)* (Mei et al., 2007), *Joint Sentiment/Topic model (JST)* (He et al., 2011) and *Aspect and Sentiment Unification Model (ASUM)* (Jo and Oh, 2011) have been proposed in the literature to jointly perform sentiment classification and topic modeling. The main advantage of the joint modeling of sentiments and aspects comes from its ability to reciprocally reduce the noise of both tasks. However, these techniques consider textual information only. We will see that including the network leads to the ability of accessing additional information when words are not sufficient (e. g., *"I cannot wait to watch that movie!!"*) or their sentiments are contradictory with the overall message polarity (e. g., *"I'm unqualified to understand the potential of Android."*). In the following, an unsupervised probabilistic model called Networked Aspect-Sentiment model (NAS) is presented. It incorporates approval relations to perform the sentiment classification and aspect extraction tasks simultaneously.

Hashtags usually identify the overall topic of a text message and not its sub-topics

The joint modeling of sentiments and aspects reciprocally reduces the noise of both tasks

7.2 THE CLASSIFICATION MODEL

In this section, the generative process of Networked Aspect-Sentiment model (NAS) is presented and how the H-DAG (ϕ) is used in the model inference is shown.

Generative Process: formally, the NAS generative process is defined as follows (we follow the notations in Table 8):

1. For each message m,
   a) Draw a distribution $\pi_m \sim \mathrm{Dirichlet}(\gamma)$ over sentiment labels.
   b) For each sentiment s in message m, draw a distribution $\theta_{m,s} \sim \mathrm{Dirichlet}(\alpha)$ over aspects.
   c) For each word w_t in message m:
      i. Choose a sentiment label $s_t \sim \pi_m$,
      ii. Choose an aspect $z_t \sim \theta_{m,s_t}$,
      iii. Choose a word w_t from the word distribution $\varphi_{z_t}^{s_t}$ over aspect z_t and sentiment label s_t, where $\varphi_{z_t}^{s_t} \sim \mathrm{Dirichlet}(\beta)$.

v_i	user
(v_i, v_j)	directed edge from user v_i to user v_j
V, E	set of users and approval edges
$M(v_j)$	set of messages of user v_j
$c_{i,j}$	normalized weight of edge (i, j)
T, S, W	number of topics, sentiments and distinct terms
m, z, s, w	message, aspect, sentiment and word
z_{-t}, s_{-t}	vector of assignments of aspects and sentiments for all the words in the corpus, except for the t-th word
π, θ, φ	multinomial distribution over sentiments, aspects and words
α, β, γ	Dirichlet prior vectors for θ, φ and π
$N_{w_t,k,l}^{-t}$	number of times that word w_t appeared in aspect k with sentiment l, except for the t-th word
$N_{k,l}^{-t}$	number of times that words are assigned sentiment l and aspect k, except for the t-th word
$N_{k,l,m}^{-t}$	number of times that word w_t is assigned sentiment l and aspect k in message m, except for the t-th word
$N_{l,m}^{-t}$	number of times that sentiment l is assigned to message m, except for the t-th word
N_m^{-t}	total number of words in the corpus, except for the t-th word

Table 8: Meaning of the NAS notation

Since some positive words, as for example 'good' and 'great', are less probable in negative expressions, and analogously, negative words as 'bad' and 'annoying' are unlikely in positive expressions, this prior information has been encoded into an asymmetric hyper-parameter β, such that its entries corresponding to general positive sentiment words have small values for negative senti-aspects, and vice-versa. Similarly, the hyper-parameters α and γ represent the prior for the aspect distribution θ and the sentiment distribution π in a message m, respectively.

Inference: the sampling distribution for a word w_t given the remaining aspects and sentiment labels can be written as $p(z_t = k, s_t = l \mid \mathbf{w}, z_{-t}, s_{-t}, \alpha, \beta, \gamma)$, where $\mathbf{w}$ is the vocabulary vector. By marginalizing out the random variables φ, θ and π by Gibbs sampling (Griffiths and Steyvers, 2004), the resulting probability can be written as:

$$p(z_t = k, s_t = l \mid \mathbf{w}, z_{-t}, s_{-t}, \alpha, \beta, \gamma) \propto \frac{N^{-t}_{w_t,k,l} + \beta}{N^{-t}_{k,l} + W\beta} \times \frac{N^{-t}_{k,l,m} + \alpha}{N^{-t}_{l,m} + T\alpha} \times \frac{N^{-t}_{l,m} + \gamma}{N^{-t}_{m} + S\gamma} \tag{44}$$

which corresponds to the traditional JST model inference (Section 3.4.2). The inclusion of Approval Network in Equation (44) leads to the following probability model:

$$p(z_t = k, s_t = l \mid \mathbf{w}, z_{-t}, s_{-t}, \alpha, \beta, \gamma, \phi) \propto \overbrace{\frac{N^{-t}_{w_t,k,l} + \beta}{N^{-t}_{k,l} + W\beta} \times \frac{N^{-t}_{k,l,m} + \alpha}{N^{-t}_{l,m} + T\alpha} \times \frac{N^{-t}_{l,m} + \gamma}{N^{-t}_{m} + S\gamma}}^{\mathbf{A}\text{ (i. e. JST)}} \times \overbrace{\frac{\sum_{(v_i,v_j)\in E} c_{i,j} \overbrace{\frac{1}{|M(v_j)|} \sum_{m' \in M(v_j)} \frac{N^{-t}_{l,m'} + \gamma}{N^{-t}_{m'} + S\gamma}}^{\mathbf{B'}}}{\sum_{(v_i,v_j)\in E} c_{i,j}}}^{\mathbf{B}\text{ (i. e. Approval Network)}} \tag{45}$$

Equation (45) combines the traditional JST model (**A**) with the contribution provided by H-DAG (**B**). The rationale behind (**B**) is to smooth JST according to the weight of the relationship that links the source user v_i and the target user v_j. In particular, JST is tuned according to the factor (**B′**) representing the average probability that the terms of messages m' of v_j (excluding term t) belong to sentiment l. This averaged probability is further regularized by weight c_{ij}. The entire smoothing process is based on a projection of weights c_{ij} between users and their corresponding messages (see Figure 30).

7.3 EXPERIMENTAL INVESTIGATION

In this section, the experimental results on real data are presented to demonstrate that the inclusion of the approval network handled through NAS outperforms the state-of-the-art baseline models for both sentiment classification and aspect extraction tasks.

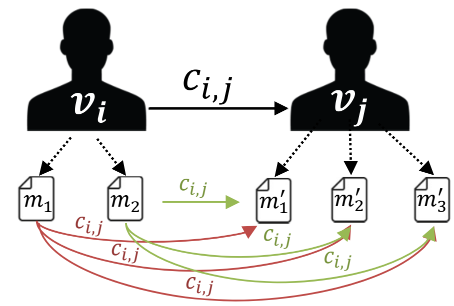


Figure 30: Projection of H-DAG: given two users connected by an edge with weight $c_{i,j}$, all the messages of the source user are connected with the messages of the target user by a projection of the weight $c_{i,j}$.

7.3.1 *Datasets and Settings*

Two datasets from Twitter have been collected by monitoring users (tweets and retweets) posting on iOS7 and "Lone Survivor". iOS7 contains positive, negative and neutral tweets, while "Lone Survivor" does not contain negative tweets.

In order to transform microblog messages to a more canonical language, URLs, mention tags (@), hashtags (#) and retweet (RT) symbols have been removed and misspelled tokens have been automatically corrected using the Google's Spell Checker API[1]. Emoticons (e. g., ':-)', ':(', ...), initialisms for emphatic expressions (e. g., 'ROFL', 'LMAOL', 'ahahah', 'eheh', ...) and onomatopoeic expressions (e. g., 'bleh', 'wow', ...) are not removed but instead are replaced by `POS_EXPRESSION`, `NEU_EXPRESSION` and `NEG_EXPRESSION`, depending on their sentiment (retrieved by lexicons a priori defined).

For a direct comparison of NAS with state-of-the-art approaches, four models have been considered. The first baseline for sentiment classification purposes is the Dictionary-based classifier (DIC) (Hu and Liu, 2004), where the overall message polarity is determined by first checking whether each term belongs to the positive, negative or neutral lexicon and then using Majority Voting. Additionally, traditional approaches able to jointly model sentiments and aspects as JST, ASUM and TSM have been compared with NAS. Posterior inference was drawn using 1000 Gibbs iterations with 100 burn-in iterations, while for TSM the Kullback-Leibler divergence measure is calculated over the sentiment coverage in order to determine the model convergence. Supported by a detailed investigation of the main aspects into the data, the number of aspects has been set $T = 10$ for iOS7, while $T = 5$ for "Lone Survivor".

1 https://code.google.com/p/google-api-spelling-java/

7.3.2 *Results on Sentiment classification*

In order to evaluate the performance of the considered models for the sentiment classification task, the classical state-of-the-art measures for classification have been employed: Precision (P), Recall (R), and F1-measure (F1). Accuracy has not been computed due to the imbalanced classes in the datasets. NAS performance has been compared with the baseline models using the macro-average F1-measure (which aggregates both Precision and Recall) over all the sentiments. Two annotators manually labeled 1,000 tweets (randomly sampled) for each dataset.

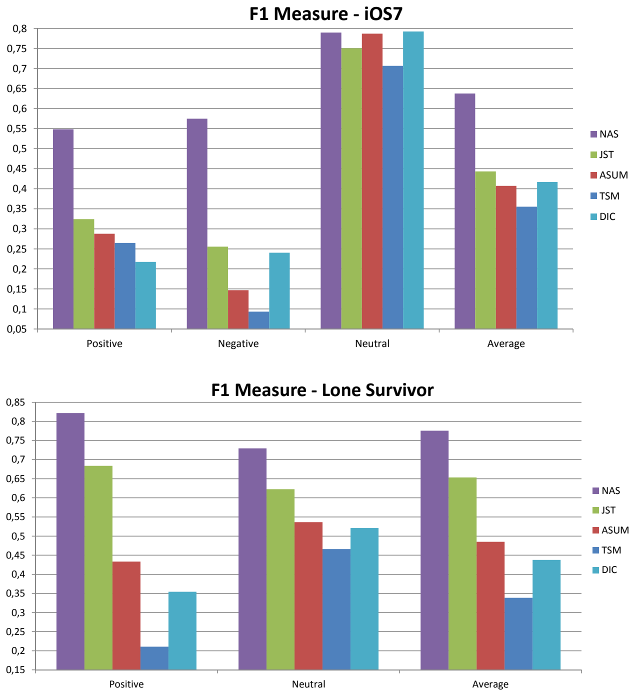


Figure 31: F1-measure on iOS7 and "Lone Survivor"

If we focus on Figure 31, we can note the outperforming results of NAS with respect to the other approaches. The first consideration relates to the direct comparison of NAS with the popular baseline DIC. As expected, DIC achieves low performance of F1-measure (0.41 on iOS7 and 0.43 on "Lone Survivor"). Conversely, NAS is able to achieve valuable improvements showing macro-average F1-measure of 0.63 on iOS7 and 0.77 on "Lone Survivor".

NAS is able to achieve the best results on iOS7, leading to a gain of 23% compared with ASUM, 19% compared with JST and 28% compared with TSM[2]. Regarding "Lone Survivor", NAS achieves a gain of 29% over ASUM, 12% over JST and 42% over TSM.

The outperforming results of NAS are due to several reasons. First, NAS overcomes the ASUM assumption stating that the sentiments of all

2 Please note that NAS corresponds to JST + the network.

the words of a given sentence have to be consistent each other. Second, NAS works well with positive, negative and neutral opinions because it is able to deal with the co-occurrence of different sentiments of seed words. For instance, consider the sentence *"I bought the **new** iPhone and the screen is very **nice**."*, where a neutral word (new) co-occurs with a positive word (nice) in the same sentence. Finally, NAS shows that including the network leads to the ability of accessing additional information when words are not sufficient or their sentiments are contradictory with the overall message polarity (i. e. NAS is capable of detecting positive tweets when none of their words are positive and vice-versa). These abilities are confirmed by some examples reported in Table 9, where tweets are correctly classified only by NAS.

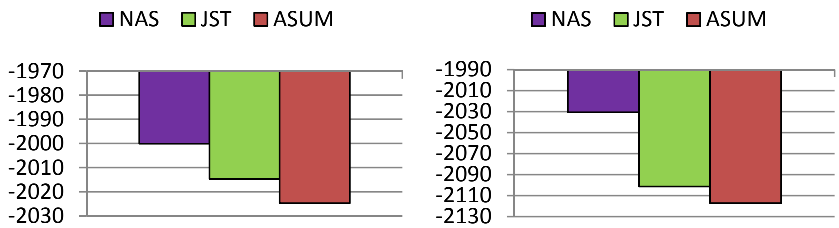
NAS	JST	ASUM	Tweet
POS	NEG	NEG	If Lone Survivor didn't change your life, you're fucking insane.
NEU	NEG	NEG	So I know how to get iOS 7 but I'm to stupid to figure out what I'm doing
POS	NEU	NEU	Lone survivor time!!
NEG	POS	NEU	Why are you thinking of switching? iOS7 looks great, just like Android POS_EXPRESSION
NEG	NEU	NEU	iOS 7 looks like a child's coloring book!!

Table 9: Examples of tweets whose sentiment is correctly captured by the network inclusion (NAS)

For example, the tweet *'iOS 7 looks like a child's coloring book!!'* is correctly classified as negative from NAS, even if it does not contain negative words. Conversely, the tweet *'If Lone Survivor didn't change your life, you're fucking insane'* as been classified by NAS as positive towards "Lone Survivor", even if it contains only neutral and negative words (i.e. fucking). In these two examples, NAS correctly classifies them as negative/positive because their authors retweet other users who emit negative/positive tweets.

7.3.3 *Results on Aspect Discovery*

In order to automatically measure the quality of the aspects extracted by NAS, JST, ASUM and TSM, the *Topic Coherence* measure (Mimno et al., 2011) has been adopted. This measure depends on the corpus without using any other resources and is computed according to the co-document frequency ratio among top topical terms. Note that, given the initial number of aspects T and sentiments S, NAS, JST and ASUM produce $T \times S$ aspects. Since TSM performs the worst in sentiment classification and it has T aspects as output, Topic Coherence has not been computed because it is sensitive to the number of aspects (Mimno et al., 2011). If we focus on Figure 32, we can note that the inclusion of the network through NAS leads to significant improvements on both datasets. ASUM achieves the lowest performance, followed by JST[2].



(a) iOS7 (b) Lone Survivor

Figure 32: Comparison of Topic Coherence

Even if Topic Coherence is a good measure to compare aspects extracted by different models, it is not able to provide a clear idea of their pros and cons. For this reason, some examples (Table 10) are shown and discussed. Although only "Battery life" and "Security" are shown, NAS improves in many other aspects. "Security" has been weakly identified for ASUM only by the words 'privacy' and 'unlocked', because the model tends to put nouns in neutral aspects, while only adjectives in positive and negative aspects, which clearly makes their characterization very difficult.

BATTERY LIFE			SECURITY		
ASUM (neu)	JST (neu)	NAS (n)	ASUM (n)	JST (n)	NAS (n)
screen	**battery**	**battery**	shit	bad	**lock**
battery	**life**	**life**	ugly	**security**	*hidden*
lock	**os**	**half**	hate	**major**	**flaw**
panoramas	**half**	bad	~~love~~	*true*	**privacy**
~~cool~~	**problem**	negexp	negexp	**flaw**	**security**
video	*unauthorized*	fast	bad	ugly	annoying
panoramic	*toy*	**problem**	horrible	negexp	scared
background	*drop*	**night**	tough	**privacy**	ugly
wallpaper	**update**	**day**	**unlocked**	annoying	negexp
lock-screen	*app*	**hours**	**privacy**	**unauthorized**	hate

Table 10: Example aspects in iOS7. ~~Strike~~/underline indicate incoherent/coherent adjectives, respectively; **bold**/*italic* indicate coherent/incoherent nouns, respectively.

Moreover, inconsistent adjectives for positive and negative aspects (e. g., 'love' for a negative aspect) and several non-related words are chosen by ASUM. Instead, JST and NAS have a good balance of adjectives and nouns for positive and negative aspects that allow us to easily characterize the aspect and simultaneously understand its perceived sentiment. The authors manually investigated the data and found out that "Battery life" is negatively perceived because of its short life and "Gaming" is positively perceived thanks to the introduction of controllers' support and kits. Unlike neutral opinions, positive and negative opinions are mostly identifiable by the presence of opinionated words (e. g., adjectives). According to this consideration, only NAS leads to the correct identification of the 'Battery Life' thanks to the ability to include opinionated words. Conversely, JST has a more clear behavior in excluding opinionated words, which generally leads the model to infer the neutral sentiment. The experimental results suggest that the introduction of the network not only improves sentiment classification, but also the ability of identifying aspects. This is because the joint modeling of sentiments and aspects has the ability to reciprocally reduce the noise of both tasks.

8 CONCLUSION AND FUTURE WORK

"In literature and in life we ultimately pursue, not conclusions, but beginnings."
— Sam Tanenhaus, Literature Unbound

Although linguistics and Natural Language Processing (NLP) have a long history, Sentiment Analysis has grown to be one of the most active research areas in NLP only since early 2000, when it has become a very active research area for different reasons. First, it has a wide array of applications, almost in every domain. The industry surrounding SA has also flourished due to the proliferation of commercial applications, which provides a strong motivation for research. Second, it offers many challenging research problems, which had never been studied before. Third, for the first time in human history, we now have a huge volume of opinionated data recorded and easily accessible in digital forms on the Web. Without it, a lot of research would not have been possible. In particular, the inception and rapid growth of the field in the last years coincide with those of the Social Media on the Web. However, considering the differences between well-formed documents (e. g., news or reviews) and Social Networks where opinions are full of hasty spelling and unconventional grammatical quirks, the strategies available in the current state of the art are no longer effective for mining opinions in this new and challenging environment. Although most of the works regarding polarity classification usually consider text as unique information to infer sentiment, Social Networks, the today's largest and richest sources of information, are actually networked environments. A representation of real world data where instances are considered as homogeneous, independent and identically distributed (i.i.d.) lead us to a substantial loss of information and to the introduction of a statistical bias. For this reason, starting from the classical state-of-the-art methodologies where only text is used to infer the polarity of social media messages, this thesis presented novel probabilistic relational models both on document and aspect level which integrate this structural information to improve classification performance. The experimental investigations revealed that incorporating probabilistic networks can lead to statistically significant improvements over the performance of complex supervised classifiers based only on textual features. The combination of content and relationships is a core task of the recent literature on Sentiment Analysis, where friendships are usually investigated to model the principle of homophily, which states that a contact among similar people occurs at a higher rate than among dissimilar people. However, paired with the assumption of homophily, constructuralism explains how social relationships evolve via dynamic and continuous interactions as the knowledge and behavior that two actors share increase. Considering the similarity among users on the basis of constructuralism appears to be a much more powerful force than interpersonal influence within the friendship network. For this reason, this thesis proposed Approval Network as a novel graph representation to jointly model homophily and constructuralism, which is intended to better represent the contagion on social networks. Two case studies which employ Approval Network confirm that the inclusion of relationships in predictive models, as suggested in other studies, leads to improve recognition performance when dealing with non-propositional environments.

There are a number of interesting ongoing researches which arise from the ideas presented in this thesis. Each of them are discussed below, according to the main contributions presented in this thesis:

1. **Approval Network** (Chapter 4): the main assumption made on Approval Network states that two connected users are likely to share the same sentiment on a given topic. For future work, this assumption can be relaxed considering that a neutral user could also be approved by negative or positive users and vice-versa.

2. **A content-based approach for polarity classification at document-level: Bayesian Model Averaging** (Chapter 5): the experimental results show that the proposed solution is particularly effective and efficient, thanks to its ability to define a strategic combination of different classifiers through an accurate and computationally efficient heuristic. However, an increasing number of classifiers to be enclosed in the ensemble together with large dataset open to deeper considerations in terms of complexity. The selection of the initial ensemble should consider the different complexities of each single learner and inference algorithm, leading to a reasonable trade-off between their contribution in terms of accuracy and the related computational time. A further ongoing research is related to the development of a hierarchical ensemble framework where the discrimination between "objective" and "subjective" is firstly addressed, to then approach the polarity classification of subjective expressions, possibly considering a wider range of labels.

3. **Exploiting network information for polarity classification: document and user-level** (Chapter 6): due to the lack of datasets which comprise the different characteristics that the proposed model needs (user-user connections, labeled posts, labeled users,

etc.), the experiments have been performed on a relatively small dataset. An ongoing research could certainly regard with the experimental investigation on larger datasets. Moreover, in addition to positive and negative, the neutral sentiment could also be considered.

4. **Aspect-Sentiment modeling using network** (Chapter 7): given a message, the proposed solution aims at extracting its main topic and classifying its sentiment. However, the extraction of multiple aspects belonging to the same message and their sentiment classification could be investigated.

A BAGGING COMBINATION RULES VS BASELINE CLASSIFIERS

A.1 MOVIEDATA DATASET

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	63.13	-	-	-	-
	R_-	73.26	-	-	-	-
	$F1_-$	67.81	-	-	-	-
	P_+	68.11	-	-	-	-
	R_+	57.16	-	-	-	-
	$F1_+$	62.14	-	-	-	-
	Acc	65.21	-	-	-	-
NB	P_-	78.03	77.92	77.16	77.72	77.66
	R_-	78.56	78.46	78.05	78.31	78.29
	$F1_-$	78.28	78.18	77.59	78.01	77.97
	P_+	78.42	78.32	77.80	78.15	78.12
	R_+	77.86	77.75	76.89	77.54	77.47
	$F1_+$	78.13	78.02	77.34	77.84	77.79
	Acc	**78.21**	78.11	77.47	77.93	77.88
ME	P_-	79.09	79.35	78.80	79.56	79.61
	R_-	77.92	76.55	76.17	76.42	76.45
	$F1_-$	78.47	77.90	77.44	77.93	77.97
	P_+	78.25	77.38	76.96	77.34	77.38
	R_+	79.34	80.04	79.47	80.32	80.38
	$F1_+$	78.77	78.66	78.18	78.78	78.83
	Acc	**78.63**	78.29	77.82	78.37	78.41
SVM	P_-	75.57	76.45	75.85	76.38	76.43
	R_-	74.95	74.86	75.42	75.50	75.52
	$F1_-$	75.23	75.63	75.61	75.92	75.95
	P_+	75.20	75.40	75.57	75.80	75.83
	R_+	75.74	76.92	75.93	76.62	76.68
	$F1_+$	75.44	76.14	75.72	76.19	76.23
	Acc	75.35	75.89	75.68	76.06	**76.10**
CRF	P_-	76.74	76.99	76.39	76.74	76.75
	R_-	74.47	74.88	75.50	75.98	75.82
	$F1_-$	75.57	75.91	75.93	76.35	76.26
	P_+	75.18	75.55	75.80	76.22	76.12
	R_+	77.37	77.60	76.64	76.94	77.00
	$F1_+$	76.25	76.55	76.20	76.57	76.54
	Acc	75.92	76.24	76.07	**76.46**	76.41

Table 11: Baseline classifier and Bagging performance (%) on MovieData with Boolean weighting schema

A.2 PRODUCTDATA

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	75.60	-	-	-	-
	R_-	75.45	-	-	-	-
	$F1_-$	75.49	-	-	-	-
	P_+	64.66	-	-	-	-
	R_+	64.78	-	-	-	-
	$F1_+$	64.65	-	-	-	-
	Acc	71.07	-	-	-	-
NB	P_-	70.59	70.64	69.67	70.12	70.03
	R_-	86.59	87.58	86.59	87.42	87.58
	$F1_-$	77.54	77.94	77.00	77.58	77.59
	P_+	70.88	72.13	69.87	71.40	71.34
	R_+	46.96	46.30	44.78	45.22	44.89
	$F1_+$	55.46	55.19	53.58	54.25	54.01
	Acc	70.31	**70.63**	69.42	70.09	70.04
ME	P_-	74.80	74.12	73.63	74.28	74.41
	R_-	75.08	76.21	75.53	76.29	75.98
	$F1_-$	74.72	74.92	74.40	75.04	74.95
	P_+	63.51	63.86	62.78	63.94	63.86
	R_+	62.72	60.76	60.22	60.98	61.41
	$F1_+$	62.72	61.83	61.15	62.00	62.18
	Acc	**70.00**	69.87	69.24	**70.00**	**70.00**
SVM	P_-	72.88	72.73	71.91	71.74	71.81
	R_-	73.71	75.68	78.33	77.88	78.11
	$F1_-$	73.22	74.07	74.88	74.60	74.73
	P_+	61.67	63.29	64.57	64.18	64.44
	R_+	60.43	59.13	55.87	55.87	55.87
	$F1_+$	60.88	60.87	59.62	59.48	59.58
	Acc	68.26	68.88	**69.11**	68.84	68.97
CRF	P_-	70.57	69.84	69.71	69.89	70.08
	R_-	74.02	76.06	77.12	77.50	77.95
	$F1_-$	72.16	72.74	73.16	73.42	73.72
	P_+	59.63	60.55	61.32	61.78	62.26
	R_+	55.33	52.61	51.74	51.85	51.96
	$F1_+$	57.23	56.11	55.94	56.13	56.39
	Acc	66.34	66.43	66.70	66.96	**67.28**

Table 12: Baseline classifier and Bagging performance (%) on ProductData with Boolean weighting schema

A.3 PRODUCTDATAMD (BOOKS)

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	73.94	-	-	-	-
	R_-	56.40	-	-	-	-
	$F1_-$	63.90	-	-	-	-
	P_+	64.70	-	-	-	-
	R_+	79.90	-	-	-	-
	$F1_+$	71.46	-	-	-	-
	Acc	68.15	-	-	-	-
NB	P_-	78.52	78.59	79.37	78.33	78.71
	R_-	84.40	84.20	84.00	84.00	84.30
	$F1_-$	81.21	81.13	81.45	80.92	81.24
	P_+	83.27	83.15	83.27	82.95	83.34
	R_+	76.50	76.60	77.80	76.40	76.80
	$F1_+$	79.53	79.51	80.21	79.32	79.70
	Acc	80.45	80.40	**80.90**	80.20	80.55
ME	P_-	79.44	80.29	80.17	80.64	80.89
	R_-	83.60	84.00	83.70	83.50	83.60
	$F1_-$	81.20	81.95	81.74	81.89	82.06
	P_+	82.84	83.36	83.07	83.08	83.22
	R_+	77.60	79.00	78.90	79.60	79.90
	$F1_+$	79.79	80.93	80.74	81.10	81.33
	Acc	80.60	81.50	81.30	81.55	**81.75**
SVM	P_-	76.74	77.70	77.80	77.89	77.68
	R_-	77.60	78.30	77.40	78.40	78.60
	$F1_-$	77.06	77.88	77.46	78.01	78.01
	P_+	77.40	78.14	77.51	78.29	78.41
	R_+	76.20	77.20	77.50	77.40	77.10
	$F1_+$	76.66	77.53	77.35	77.69	77.59
	Acc	76.90	77.75	77.45	**77.90**	77.85
CRF	P_-	78.20	78.46	77.76	78.15	78.01
	R_-	77.20	77.50	77.50	78.10	78.40
	$F1_-$	77.59	77.82	77.46	77.95	78.03
	P_+	77.51	77.77	77.59	78.14	78.27
	R_+	78.20	78.30	77.40	77.70	77.40
	$F1_+$	77.74	77.87	77.32	77.73	77.65
	Acc	77.70	**77.90**	77.45	**77.90**	**77.90**

Table 13: Baseline classifier and Bagging performance (%) on ProductDataMD (books) with Boolean weighting schema

A.4 PRODUCTDATAMD (MUSIC)

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	80.31	-	-	-	-
	R_-	46.00	-	-	-	-
	$F1_-$	58.22	-	-	-	-
	P_+	62.28	-	-	-	-
	R_+	88.70	-	-	-	-
	$F1_+$	73.12	-	-	-	-
	Acc	67.35	-	-	-	-
NB	P_-	79.27	78.60	79.76	79.20	79.50
	R_-	81.30	80.10	81.40	80.70	81.00
	$F1_-$	80.21	79.26	80.50	79.86	80.15
	P_+	80.97	79.86	81.18	80.46	80.79
	R_+	78.70	78.10	79.30	78.70	79.00
	$F1_+$	79.75	78.89	80.16	79.49	79.80
	Acc	80.00	79.10	**80.35**	79.70	80.00
ME	P_-	81.55	81.16	81.24	81.05	81.15
	R_-	79.00	76.80	78.40	77.10	77.80
	$F1_-$	80.10	78.85	79.68	78.94	79.34
	P_+	79.87	78.10	79.26	78.30	78.81
	R_+	81.90	82.10	81.70	81.90	81.80
	$F1_+$	80.72	79.99	80.35	79.99	80.19
	Acc	**80.45**	79.45	80.05	79.50	79.80
SVM	P_-	74.88	76.61	75.28	76.60	76.51
	R_-	74.80	75.50	74.60	76.20	76.10
	$F1_-$	74.76	75.98	74.87	76.33	76.24
	P_+	74.93	75.89	74.92	76.40	76.29
	R_+	74.80	76.80	75.40	76.60	76.50
	$F1_+$	74.79	76.27	75.09	76.43	76.32
	Acc	74.80	76.15	75.00	**76.40**	76.30
CRF	P_-	76.75	75.84	74.57	75.18	75.51
	R_-	75.30	75.00	74.80	76.30	76.80
	$F1_-$	75.94	75.36	74.62	75.66	76.07
	P_+	75.86	75.41	74.79	76.05	76.51
	R_+	77.10	76.10	74.40	74.70	75.00
	$F1_+$	76.41	75.70	74.53	75.29	75.67
	Acc	**76.20**	75.55	74.60	75.50	75.90

Table 14: Baseline classifier and Bagging performance (%) on ProductDataMD (music) with Boolean weighting schema

A.5 GOLD STANDARD MOVIE

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	37.46				
	R_-	83.33	-	-	-	-
	$F1_-$	50.90	-	-	-	-
	P_+	95.26	-	-	-	-
	R_+	68.10	-	-	-	-
	$F1_+$	78.92	-	-	-	-
	Acc	70.78	-	-	-	
NB	P_-	41.24	44.50	42.70	46.95	46.54
	R_-	26.67	28.89	27.78	28.89	28.89
	$F1_-$	31.94	34.20	33.21	35.12	34.92
	P_+	85.50	85.81	85.68	85.89	85.85
	R_+	92.38	91.90	92.38	92.62	92.38
	$F1_+$	88.77	88.71	88.87	89.10	88.96
	Acc	80.78	80.78	80.98	**81.37**	81.18
ME	P_-	48.10	47.76	46.82	50.42	50.06
	R_-	35.56	30.00	33.33	30.00	31.11
	$F1_-$	39.91	36.38	38.00	36.94	37.73
	P_+	87.10	86.23	86.81	86.38	86.54
	R_+	92.38	93.57	93.10	94.52	94.29
	$F1_+$	89.61	89.72	89.79	90.23	90.22
	Acc	82.35	82.35	82.55	**83.14**	**83.14**
SVM	P_-	58.54	65.42	64.31	65.25	62.58
	R_-	51.11	44.44	33.33	34.44	33.33
	$F1_-$	52.64	50.19	42.10	42.93	41.46
	P_+	90.07	89.05	87.18	87.39	87.17
	R_+	92.62	95.00	96.19	96.19	95.95
	$F1_+$	91.22	91.85	91.42	91.52	91.30
	Acc	85.29	**86.08**	85.10	85.29	84.90
CRF	P_-	59.17	48.83	32.50	48.33	48.33
	R_-	25.56	13.33	7.78	11.11	11.11
	$F1_-$	34.59	20.01	12.17	17.28	17.28
	P_+	85.94	83.90	83.32	83.77	83.80
	R_+	96.90	96.67	98.57	98.10	98.33
	$F1_+$	91.05	89.82	90.29	90.35	90.47
	Acc	**84.31**	81.96	82.55	82.75	82.94

Table 15: Baseline classifier and Bagging performance (%) on Gold Standard Movie with Boolean weighting schema

A.6 GOLD STANDARD PERSON

		BASELINE	MV	MAX	MEAN	PROD
DIC	P_-	53.97	-	-	-	-
	R_-	82.00	-	-	-	-
	$F1_-$	64.81	-	-	-	-
	P_+	92.62	-	-	-	-
	R_+	75.52	-	-	-	-
	$F1_+$	83.04	-	-	-	-
	Acc	77.18	-	-	-	-
NB	P_-	62.51	62.68	62.74	63.33	62.96
	R_-	56.00	54.00	54.00	56.00	56.00
	$F1_-$	57.78	56.90	56.61	58.08	57.78
	P_+	85.37	84.93	85.04	85.38	85.47
	R_+	87.24	88.28	88.28	87.59	87.59
	$F1_+$	86.14	86.46	86.48	86.35	86.37
	Acc	79.23	**79.49**	**79.49**	**79.49**	**79.49**
ME	P_-	65.83	65.56	63.83	64.60	63.39
	R_-	49.00	46.00	47.00	45.00	44.00
	$F1_-$	54.75	53.37	53.05	51.86	50.81
	P_+	83.97	83.07	83.34	82.92	82.68
	R_+	90.69	91.03	90.34	91.03	91.03
	$F1_+$	87.09	86.82	86.60	86.70	86.56
	Acc	**80.00**	79.49	79.23	79.23	78.97
SVM	P_-	55.43	53.55	57.48	56.37	59.17
	R_-	43.00	32.00	31.00	28.00	31.00
	$F1_-$	47.50	39.09	39.18	36.14	39.64
	P_+	81.69	79.15	79.40	78.77	79.53
	R_+	87.24	88.97	91.38	91.72	92.07
	$F1_+$	84.25	83.70	84.89	84.69	85.28
	Acc	75.90	74.36	75.92	75.38	**76.41**
CRF	P_-	67.51	70.36	65.83	67.23	67.70
	R_-	38.00	34.00	26.00	30.00	31.00
	$F1_-$	47.68	44.73	35.01	39.94	40.82
	P_+	81.47	80.73	79.18	80.00	80.25
	R_+	93.45	94.83	95.86	95.52	95.52
	$F1_+$	87.00	87.16	86.63	86.99	87.13
	Acc	**79.23**	79.22	77.95	78.72	78.97

Table 16: Baseline classifier and Bagging performance (%) on Gold Standard Person with Boolean weighting schema

ACKNOWLEDGEMENTS

ENGLISH:

Within "The Divine Comedy", it is represented a fundamental educational journey that Dante, together with the reader, is called to make. For this reason, I chose to accompany my acknowledgments with some comparisons between people who have been essential in achieving this goal and the characters of the Poem.

Betta, you have accompanied me gradually towards all objectives, helping me in writing papers and presentations. We spent entire evenings in the university. You have encouraged and stimulated me when the discomfort has prevailed. You taught me to produce high-quality scientific papers and create clear and complete presentations. You taught me to identify problems and find appropriate solutions. You taught me to formalize concepts. You taught me pretty much everything.

Prof., you have been my guide for more than five years. My respect and my admiration has always been huge. I really appreciate your honesty, your professionalism, skills and, last but not least, your ability to give rise to a peaceful and pleasant work environment. I appreciate how much you have done for me. I will never forget it.

In my "comparisons game", both of you are Virgilio, who has a particular symbolic value for the entire Dante's journey: he has the duty to provide him the foundation, making him able to receive teachings from Stazio, his second guide in this long journey. In hell, Virgilio performs its function efficiently, averting the pilgrim in principle by a hard road that leads to Monte Purgatorio, holding off the guardians of hell and finally exhorting or admonishing his protected. With suitable abstractions, it seems to me an apt comparison.

Prof. Liu, it has been a great honor to work with you. Just consider that I approached the "Sentiment Analysis" research filed by reading and studying your papers. In addition to having met a guru in this field, I met a humanly wonderful person. Thanks a lot for everything you taught me.

Finally my family, which means Mom, Dad, Clara, Cristina and Alessio. I love you more than anything else and, although I am sometimes surly, sullen and unbearable, I would do everything for you, even if it is hard to think to be able to do as much as you have done for me. You always supported me in my personal choices as well as in their realization, through an immense love and lots of sage advice.

At the sight of Dante, Caronte shouts him to get away from there, because others will ferry him to the kingdom of darkness. At this point, an earthquake, accompanied by a blinding lightning, makes the earth tremble so violently that Dante is terrified and lose consciousness, falling to the ground. Although Caronte did not want him on his boat, his words are a prophecy of salvation. I chose this comparison because you have been able, in the past, present, and I am sure you will be in the future, to guide me towards "the right", keeping me away from the mistakes that I would have easily done. Your advice has always been valuable, since I was young and I did not want to study, when I grew up and began to walk the tortuous path of life and now that I am beginning a new phase of my life. You have shown me love avoiding the search of the qualities that I do not own, excusing me all the times I made mistakes because of this. Instead, you have identified and appreciated the qualities that I really own, loving me. Thank you so much for this, that I consider pure love, a love that one day I hope that I will be able to pass to my children.

ITALIAN:

All'interno de "La Divina Commedia", è rappresentato un percorso educativo fondamentale che Dante, insieme al lettore, è portato a compiere. Proprio per questo motivo, ho scelto di accompagnare i miei ringraziamenti con alcuni paragoni tra le persone che sono state fondamentali per il raggiungimento di questo obiettivo e i personaggi del Poema.

Betta, mi hai accompagnato pian piano verso tutti gli obiettivi, affiancandomi durante la scrittura di articoli e presentazioni. Abbiamo passato serate intere in università. Mi hai incoraggiato e stimolato quando lo sconforto mi ha pervaso. Mi hai insegnato a produrre articoli scientifici di alta qualità e a creare presentazioni chiare e complete. Mi hai insegnato a individuare i problemi e trovare soluzioni appropriate. Mi hai insegnato a formalizzare concetti. Mi hai insegnato praticamente tutto.

Prof., è stata la mia guida per più di cinque anni. Il mio rispetto e la mia ammirazione sono sempre stati enormi. Apprezzo molto la sua onestà, la sua professionalità, le competenze e, ultimo ma non meno importante, la sua capacità di creare un ambiente di lavoro sereno e piacevole. Apprezzo quanto si sia spesa per me e questo non lo dimenticherò mai.

Nel mio "gioco dei confronti", voi siete Virgilio, il quale assume una particolare valenza simbolica per l'intero viaggio di Dante: egli ha il compito di fornirgli le basi rendendolo in grado di ricevere in seguito gli insegnamenti da Stazio, sua seconda guida in questo lungo viaggio. Nell'Inferno, Virgilio assolve il suo compito con efficienza, distogliendo in principio il pellegrino da un arduo cammino, che conduce fino al monte purgatoriale, tenendo a bada i guardiani infernali e infine esortando o ammonendo il proprio protetto. Con opportune astrazioni, mi sembra un paragone azzeccato.

Prof. Liu, è stato un grande onore lavorare con lei. Basti pensare che mi sono avvicinato alla “Sentiment Analysis” leggendo e studiando i suoi articoli. Oltre ad aver incontrato un guru in questo campo, ho incontrato una persona umanamente meravigliosa. Grazie mille per tutto quello che mi ha insegnato.

Ed infine la mia famiglia, il che vuol dire Mamma, Papà, Clara, Cristina e Alessio. Vi amo più di qualsiasi altra cosa e, nonostante a volte sia burbero, scontroso e insopportabile, farei qualsiasi cosa per voi, anche se è difficile pensare di poter fare quanto voi avete fatto per me. Mi avete sempre appoggiato nelle scelte e supportato nella loro realizzazione, attraverso un amore smisurato e tanti saggi consigli.

Alla vista di Dante, Caronte gli grida di allontanarsi di lì, perché altri lo traghetteranno verso il regno delle tenebre. A questo punto, un terremoto, accompagnato da un fulmine accecante, fa tremare la terra così violentemente che Dante ne è terrorizzato e perde i sensi, cadendo a terra. Seppur Caronte non l’abbia voluto sulla sua barca, le sue parole sono una profezia di salvezza. Ho scelto questo paragone perché siete stati capaci, nel passato, nel presente e sono sicuro lo sarete anche nel futuro, di guidarmi verso “il giusto”, tenendomi lontano dagli errori che facilmente avrei commesso. I vostri consigli sono stati sempre preziosi, fin da quando ero piccolo e non volevo studiare, quando sono cresciuto e iniziavo a percorrere il tortuoso percorso della vita ed ora che inizio una nuova fase della mia vita. Mi avete dimostrato amore evitando di cercare in me le qualità che non possiedo, perdonandomi tutte le volte che ho commesso degli errori a causa di questo. Avete invece individuato e apprezzato quelle qualità che realmente possiedo, volendomi bene. Vi ringrazio tanto per questo, che considero essere amore allo stato puro, un amore che spero riuscirò un giorno a trasmettere immutato ai miei figli.

BIBLIOGRAPHY

Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web: Research and Applications - Volume Part II*, ESWC'11, pages 375–389. Springer-Verlag.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211 – 230.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. (2001). Search in power-law networks. *Phys. Rev. E*, 64.

Aggarwal, C. C. (2011). *Social Network Data Analytics*. Springer Publishing Company, Incorporated, 1st edition.

Ahmed, N. K., Berchmans, F., Neville, J., and Kompella, R. (2010). Time-based sampling of social network activity graphs. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 1–9. ACM.

Aristotle (1934). Rhetoric. nichomachean ethics. In *Aristotle*, volume 23, page 1371. Rackman transl. Cambridge: Harvard Univ. Press.

Baldwin, T. and Lui, M. (2010). Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 229–237. Association for Computational Linguistics.

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2004). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6:2005.

Bansal, M., Cardie, C., and Lee, L. (2008). The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *COLING (Posters)*, pages 15–18.

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44. Association for Computational Linguistics.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.*, 36(1-2):105–139.

Beineke, P., Hastie, T., Manning, C., and Vaithyanathan, S. (2004). Exploring sentiment summarization. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003a). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, pages 187–205.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100. ACM.

Boiy, E. and Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558.

Bott, H. (1928). Observation of play activities in a nursery school. *Genet. Psychol. Monogr*, 4:44–88.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10. IEEE Computer Society.

Carley, K. M. (1991). A theory of group stability. *American Sociological Review*, 56(3):331–354.

Carley, K. M. and Hill, V. (2001). Structural change and learning within organizations. *Dynamics of organizations: Computational modeling and organizational theories*, pages 63–92.

Carrington, P. J., Scott, J., and Wasserman, S. (2005). *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Carter, S., Weerkamp, W., and Tsagkias, M. (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Lang. Resour. Eval.*, 47(1):195–215.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18. ACM.

Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, volume 27 of *SIGMOD '98*, pages 307–318. ACM.

Chang, J. and Blei, D. M. (2009). Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, pages 81–88.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks. *Annals of Applied Statistics*.

Chen, L., Wang, W., Nagarajan, M., Wang, S., and Sheth, A. P. (2012). Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Cheong, M. and Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, 13(1):45–59.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cobb, N. K., Mays, D., and Graham, A. L. (2013). Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. *J Natl Cancer Inst Monogr*, 47:224–230.

Cohen, J. M. (1977). Sources of peer group homogeneity. *Sociology of Education*, 50(4):227–241.

Cohn, D. and Chang, H. (2000). Learning to probabilistically identify authoritative documents. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 167–174. Morgan Kaufmann Publishers Inc.

Cohn, D. and Hofmann, T. (2001). The missing link-a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, 13:430–436.

Corbett, E. P. J. (1971). *Classical Rhetoric for the Modern Student*. Oxford University Press, 2nd ed. edition.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *ML*, 20(3):273–297.

Crane, R. and McDowell, L. K. (2011). Evaluating markov logic networks for collective classification. *Proceedings of the 9th MLG Workshop at the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Cristianini, N. and Shawe-Taylor, J. (2000). *Support Vector Machines*. Cambridge: Cambridge University Press.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528. ACM.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116. Association for Computational Linguistics.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Deng, H., Han, J., Li, H., Ji, H., Wang, H., and Lu, Y. (2013). Exploring and inferring user-user pseudo-friendship for sentiment analysis with heterogeneous networks. In *SIAM International Conference on Data Mining (SDM'2013)*, pages 378–386.

Dietterich, T. G. (2002). Ensemble learning. In *The Handbook of Brain Theory and Neural Networks*, pages 405–508. Mit Pr.

Domingos, P. and Richardson, M. (2004). Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML-2004 workshop on statistical relational learning and its connections to other fields*, pages 49–54.

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5220–5227.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Fan, W., Chu, F., Wang, H., and Yu, P. S. (2002). Pruning and dynamic scheduling of cost-sensitive ensembles. In *Eighteenth National Conference on Artificial Intelligence*, pages 146–151. American Association for Artificial Intelligence.

Farris, P. W., Bendle, N. T., Pfeifer, P. E., and Reibstein, D. J. (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Wharton School Publishing, 2nd edition.

Fersini, E. and Messina, E. (2013). Web page classification through probabilistic relational models. *International Journal of Pattern Recognition and Artificial Intelligence*.

Fersini, E., Messina, E., and Archetti, F. (2009). Probabilistic relational models with relational uncertainty: An early study in web page classification. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '09, pages 139–142. IEEE Computer Society.

Fersini, E., Messina, E., and Archetti, F. (2010a). A probabilistic relational approach for web document clustering. *Information Processing and Management*, 46(2):117–130.

Fersini, E., Messina, E., and Archetti, F. (2010b). Web page classification: A probabilistic model with relational uncertainty. In *Proceedings of the Computational Intelligence for Knowledge-based Systems Design, and 13th International Conference on Information Processing and Management of Uncertainty*, IPMU'10, pages 109–118. Springer-Verlag.

Fersini, E., Messina, E., and Pozzi, F. A. (2014a). Expressive signals in social media languages to improve polarity detection. *Information Processing & Management, IP&M 2014. Special Issue on Sentiment Analysis. Under revision.*

Fersini, E., Messina, E., and Pozzi, F. A. (2014b). Sentiment analysis: Bayesian ensemble learning. *Decision Support System. In press.*

Fersini, E., Messina, E., and Pozzi, F. A. (2014c). Subjectivity, polarity and irony detection: A multi-layer approach. In *CLIC-it 2014*.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 1300–1309. Springer-Verlag.

Gallagher, B., Tong, H., Eliassi-Rad, T., and Faloutsos, C. (2008). Using ghost edges for classification in sparsely labeled networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 256–264. ACM.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, pages 121–132. Springer-Verlag.

Gatti, M., Cavalin, P., Neto, S., Pinhanez, C., dos Santos, C., Gribel, D., and Appel, A. (2014). Large-scale multi-agent-based modeling and simulation of microblogging-based online social network. In *Multi-Agent-Based Simulation XIV*, Lecture Notes in Computer Science, pages 17–33. Springer Berlin Heidelberg.

Getoor, L., Friedman, N., Koller, D., and Taskar, B. (2003). Learning probabilistic models of link structure. *J. Mach. Learn. Res.*, 3:679–707.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 211–220. ACM.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford.

Goldberg, A. B., Zhu, X., and Wright, S. J. (2007). Dissimilarity in graph-based semi-supervised classification. In *AISTATS*, volume 2, pages 155–162.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586. Association for Computational Linguistics.

Gopal, R., Marsden, J. R., and Vanthienen, J. (2011). Information mining - reflections on recent advancements and the road ahead in data, text, and media mining. *Decision Support Systems*, 51(4):727–731.

Gouws, S., Metzler, D., Cai, C., and Hovy, E. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29. Association for Computational Linguistics.

Greene, S. C. (2007). *Spin: lexical semantics, transitivity, and the identification of implicit sentiment*. ProQuest.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Hampton, K. N. and Wellman, B. (2000). Examining community in the digital neighborhood: Early results from canada's wired suburb. In *Digital Cities*, volume 1765 of *Lecture Notes in Computer Science*, pages 194–208. Springer Berlin Heidelberg.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378. Association for Computational Linguistics.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In *SocialCom*, pages 357–364.

Hatzivassiloglou, V. and Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 299–305. Association for Computational Linguistics.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):784–796.

He, Y., Lin, C., and Alani, H. (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 123–131. Association for Computational Linguistics.

Hirshman, B. R., Charles, J., and Carley, K. M. (2011). Leaving us in tiers: Can homophily be used to generate tiering effects? *Comput. Math. Organ. Theory*, 17(4):318–343.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88. ACM.

Hopkins, D. and King, G. (2007). Extracting systematic social science meaning from text. *Manuscript available at http://gking. harvard. edu/-files/words. pdf*.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177. ACM.

Hu, X., Tang, L., Tang, J., and Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 537–546. ACM.

Hubbard, R. M. (1929). A method of studying spontaneous group formation. *Some New Techniques for Studying Social Behavior*, pages 76–85.

Huston, T. L. and Levinger, G. (1978). Interpersonal attraction and relationships. *Annual Review of Psychology*, 29:115–156.

Jeh, G. and Widom, J. (2002). Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543. ACM.

Jensen, D., Neville, J., and Gallagher, B. (2004). Why collective inference improves relational classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 593–598. ACM.

Jin, X., Li, Y., Mah, T., and Tong, J. (2007). Sensitive webpage classification for content advertising. In *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '07, pages 28–33. ACM.

Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 244–251. ACM.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824. ACM.

Joachims, T. (1999). Making large-scale support vector machine learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 169–184. MIT Press.

Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: Motives and use of facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1027–1036. ACM.

Joseph, K., Morgan, G. P., Martin, M. K., and Carley, K. M. (2013). On the coevolution of stereotype, culture, and social relationships: An agent-based model. *Social Science Computer Review*.

Kahanda, I. and Neville, J. (2009). Using transactional information to predict link strength in online social networks. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM*.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. *The American Journal of Sociology*, 84(2):427–436.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68.

Kaschesky, M. and Riedl, R. (2011). Tracing opinion-formation on political issues on the internet: A model and methodology for qualitative analysis and results. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

Kleinberg, J. M. (2006). Complex networks and decentralized search algorithms. In *In Proceedings of the International Congress of Mathematicians (ICM*, volume 3, pages 1019–1044.

Kok, S. and Domingos, P. (2009). Learning markov logic network structure via hypergraph lifting. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 505–512. ACM.

Kok, S. and Domingos, P. (2010). Learning markov logic networks using structural motifs. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 551–558. Omnipress.

Kou, Z. and Cohen, W. W. (2007). Stacked graphical models for efficient inference in markov random fields. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, pages 533–538.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc.

Landwehr, N., Kersting, K., and Raedt, L. D. (2005). nFOIL: Integrating naïve bayes and foil. In *AAAI'05*, pages 795–800.

Landwehr, N., Passerini, A., Raedt, L., and Frasconi, P. (2010). Fast learning of relational kernels. *Mach. Learn.*, 78(3):305–342.

Lazarsfeld, P. F. and Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society*, pages 18–66.

Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924. ACM.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. M. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 641–650. ACM.

Liben-Nowell, D. and Kleinberg, J. M. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559. ACM.

Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 243–252. ACM.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384. ACM.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Long, B., Zhang, Z. M., and Yu, P. S. (2007). A probabilistic framework for relational clustering. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 470–479. ACM.

Lu, Q. and Getoor, L. (2003). Link-based classification. *In Proceedings of the 20th International Conference on Machine Learning*, pages 496–503.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150.

Macskassy, S. A. (2007). Improving learning in networked data by combining explicit and mined links. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, pages 590–595. AAAI Press.

Macskassy, S. A. and Provost, F. (2003). A simple relational classifier. In *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, pages 64–76.

Macskassy, S. A. and Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983.

Marsden, P. V. (1987). Core discussion networks of americans. *American Sociological Review*, 52(1):122–131.

Marsden, P. V. (1988). Homogeneity in confiding relations. *Soc. Networks*, 10:57–76.

Martínez-Muñoz, G. and Suárez, A. (2006). Pruning in ordered bagging ensembles. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 609–616. ACM.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naïve bayes text classification. In *AAAI-98 workshop on Learning for Text Categ.*, pages 41–48.

McCallum, A., Pal, C., Druck, G., and Wang, X. (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. In *AAAI*, pages 433–439.

McDonald, S. (1999). Exploring the process of inference generation in sarcasm: A review of normal and clinical studies. *Brain and Language*, 68(3):486–506.

McPherson, M., Mayhew, B. H., Rotolo, T., and Smith-Lovin, L. (1995). Sex and ethnic heterogeneity in face-to-face groups in public places: An ecological perspective on social interaction. *Social Forces*, 74:15–52.

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

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008a). Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110. ACM.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008b). Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110. ACM.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180. ACM.

Mendes, P. N., Passant, A., Kapanipathi, P., and Sheth, A. P. (2010). Linked open social signals. In *Proceedings of the 2010*

IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '10, pages 224–231. IEEE Computer Society.

Milgram, S. (1967). The small world problem. *Psychology Today*, 67(1):61–67.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272. Association for Computational Linguistics.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42. ACM.

Mitra, G. and Mitra, L. (2011). Applications of news analytics in finance: A review. In *The Hanbook of News Analytics in Finance*. Wiley Finance.

Moghaddam, S. and Popowich, F. (2010). Opinion polarity identification through adjectives. *arXiv preprint arXiv:1011.4623*.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 412–418. Association for Computational Linguistics.

Neville, J. and Jensen, D. (2000). Iterative classification in relational data. *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20.

Neville, J. and Jensen, D. (2005). Leveraging relational autocorrelation with latent group models. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 322–329. IEEE Computer Society.

Neville, J. and Jensen, D. (2007). Relational dependency networks. *J. Mach. Learn. Res.*, 8:653–692.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A. L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*.

Pak, A. and Paroubek, P. (2010). Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 436–439. Association for Computational Linguistics.

Pang, B. and Lee, L. (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. Association for Computational Linguistics.

Parmelee, J. H. and Bichard, S. L. (2011). *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public*. Lexington Books.

Partalas, I., Tsoumakas, G., and Vlahavas, I. (2010). An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning*, 81(3):257–282.

Plato (1968). Laws. In *Plato in Twelve Volumes*, volume 11, page 837. Bury translator. Cambridge: Harvard Univ. Press.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Popescul, A., Popescul, R., and Ungar, L. H. (2003). Structural logistic regression for link analysis. In *Proceedings of the Second International Workshop on Multi-Relational Data Mining*, pages 92–106.

Pozzi, F. A., Fersini, E., and Messina, E. (2013a). Bayesian model averaging and model selection for polarity classification. In *Proc. of the 18th International Conference on Application of Natural Language to Information Systems*, LNCS, pages 189–200.

Pozzi, F. A., Fersini, E., Messina, E., and Blanc, D. (2013b). Enhance polarity classification on social media through sentiment-based feature expansion. In *WOA@ AI* IA*, pages 78–84.

Pozzi, F. A., Maccagnola, D., Fersini, E., and Messina, E. (2013c). Enhance user-level sentiment analysis on microblogs with approval relations. In *AI*IA 2013: Advances in Artificial Intelligence*, volume 8249 of *Lecture Notes in Computer Science*, pages 133–144. Springer International Publishing.

Quinlan, J. R. (1996). Bagging, boosting, and c4.5. In *Proc. of the 13th National Conf. on AI*, pages 725–730.

Rabelo, J. C., Prudêncio, R. C., and Barros, F. A. (2012). Leveraging relationships in social networks for sentiment analysis. In *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web*, WebMedia '12, pages 181–188. ACM.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256. Association for Computational Linguistics.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682. Association for Computational Linguistics.

Rattigan, M. J. and Jensen, D. (2005). The case for anomalous link discovery. *SIGKDD Explor. Newsl.*, 7(2):41–47.

Richardson, M. and Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in neural information processing systems*, 2:1441–1448.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Mach. Learn.*, 62(1-2):107–136.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534. Association for Computational Linguistics.

Rossi, R. and Neville, J. (2010). Modeling the evolution of discussion topics and communication to improve relational classification. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 89–97. ACM.

Rossi, R. A., McDowell, L., Aha, D. W., and Neville, J. (2012). Transforming graph data for statistical relational learning. *J. Artif. Intell. Res. (JAIR)*, 45:363–441.

Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., and Merom, R. (2010). Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 233–242. ACM.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Scharl, A. and Weichselbraun, A. (2008). An automated approach to investigating the online media coverage of us presidential elections. *Journal of Information Technology & Politics*, 5(1):121–132.

Scott, J. G. (2000). *Social Network Analysis: A Handbook*. SAGE Publications Ltd.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-rad, T. (2008). Collective classification in network data. *AI Magazine*, 29(3).

Shalizi, C. R. and Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239.

Sharara, H., Getoor, L., and Norton, M. (2011). Active surveying: A probabilistic approach for identifying key opinion leaders. In *IJCAI*, pages 1485–1490.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11.

Srinivasan, A. (1999). The aleph manual. *Computing Laboratory, Oxford University*.

Su, Y., Zhang, Y., Ji, D., Wang, Y., and Wu, H. (2013). Ensemble learning for sentiment classification. In *Proceedings of the 13th Chinese Conference on Chinese Lexical Semantics*, pages 84–93. Springer-Verlag.

Sun, Y., Han, J., Gao, J., and Yu, Y. (2009). itopicmodel: Information network-integrated topic modeling. In *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 493–502.

Sutton, C. A. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in ML*, 4(4):267–373.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

Täckström, O. and McDonald, R. (2011). Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405.

Taskar, B., Abbeel, P., and Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 485–492. Morgan Kaufmann Publishers Inc.

Taskar, B., Wong, M. F., Abbeel, P., and Koller, D. (2003). Link prediction in relational data. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*. MIT Press.

Titov, I. and Mcdonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *ACL*, pages 308–316.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424. Association for Computational Linguistics.

Verbrugge, L. M. (1977). The structure of adult friendship choices. *Soc. Forces*, 56:576–597.

Verbrugge, L. M. (1983). A research note on adult friendship contact: a dyadic perspective. *Soc. Forces*, 62:78–83.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243. Association for Computational Linguistics.

Wang, G., Sun, J., Ma, J., Xu, K., and Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57(0):77–93.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 90–94.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Watts, D. J. (2004). *Six Degrees: The Science of a Connected Age*. W. W. Norton and Company.

Wellman, B. (1929). The school child's choice of companions. *J.Educ. Res*, 14:126–32.

Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 625–631. ACM.

Wick, M., Rohanimanesh, K., Culotta, A., and McCallum, A. (2009). Samplerank: Learning preferences from atomic gradients. In *NIPS Workshop on Advances in Ranking*.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253. Association for Computational Linguistics.

Windeatt, T. and Ardeshir, G. (2004). Decision tree simplification for classifier ensembles. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(05):749–776.

Xia, R., Zong, C., and Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Inf. Sci.*, 181(6):1138–1152.

Xiang, R., Neville, J., and Rogati, M. (2010). Modeling relationship strength in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 981–990. ACM.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 189–196. Association for Computational Linguistics.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136. Association for Computational Linguistics.

Yu, S. and Kak, S. (2012). A Survey of Prediction Using Social Media. *ArXiv e-prints*.

Zaki, M. J. and Meira Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zhang, L. and Liu, B. (2011). Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 575–580. Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65. Association for Computational Linguistics.