1.1 The Demand for Social Media Learning

At the present time, the web has become a vibrant and active ecosystem in which billions of individuals around the world interact, share, post, and conduct numerous daily activities. Social media is a term used to describe Internet-based platforms that encourage individuals to easily create profiles so that they can communicatively connect and publicly share information with other individuals across different domains. Social media has enhanced the formation and exchange of user-generated content regarding the concept and technology of Web 2.0 [91]. One can undoubtedly state that social media is an important foundation for on-line interactions and contents sharing [35, 144, 169], human behavior and subjectivity [42, 90], enterprise assessments [3, 29, 95], climate change [196], health and marketing approaches [78, 99], human relationship [17, 55], social media influences [7, 10, 60], social media used in organizations [184], opinions and sentiments expressions [157, 194], and many others. This means of information has provided implications of research endeavor.

The physical world is simultaneously reflected into the virtual world where Internet-based applications are built to allow the creation, exchange of user-generated content. Social media platforms such as Facebook, Twitter, Flickr, and Foursquare enable us to be connected and interact with friends or any kind of people or organizations regardless of location and time. Many organizations, individuals and even government of countries update their activities on social media. These platforms empower organizations of any size to obtain knowledge on how their targeted audience shares information. Platforms also enable the effective spreading and collection of large-scale data giving a rise to major computational challenges. Since information is ubiquitous and overwhelmingly available, social media provides us with extraordinary opportunities to understand human behavior patterns at scale. The ultimate desire of humanity is to discovery, explain and understand the world that we are living in. By understanding individuals better, we can design better computing systems that will serve them and the community.

From the research perspective, one can see that the prospect of available and cheap access to a very large amount of data creates business and academics opportunities in a diverse range of analytical and learning contexts. *Social media learning* is an umbrella concept that contains extracting, presenting, analysing, learning meaningful patterns in different types of data from social media. It is an interdisciplinary research field that encompasses techniques from computer science, machine learning, social network analysis, data mining, natural language processing (NLP), information retrieval (IR), optimization and mathematics. With all techniques combined, the virtual world of social media is represented in a computable manner in which knowledge and information diffusion are discovered and measured. A deluge of data is unprecedentedly generated and spread by people in this information sharing society, as exemplified the following numbers:

- Every day, 3 billion people access the Internet and create approximately 2.5 quintillion bytes of data 90% of which is unstructured. By the year of 2018, 50TB of data will be generated per second [189].

- Every minute, on average, around 510 comments are posted, 293,000 statuses are updated, 3.1 million likes are generated, and 136,000 photos are uploaded on Facebook [211]. Worldwide, there are 1.13 billion daily active users on average and 1.03 billion mobile daily active users on average [57].

- Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. There are 1 billion unique visits monthly to sites with embedded tweets [186].

Billions of users participate in numerous social networks and online communities. The data these users generate mandate new computational algorithms, analysis techniques, models and applications that combine different disciplines.

1.2 Business Perspective of Geolocalization

What makes raw social media data even more interesting is that its location characteristics can be exploited, e.g. each tweet or textual post can be associated with a geographical location. These geotagged data are annotated with physical coordinates, e.g. latitude and longitude pairs, in the world geodetic system format [43]. Another geotagged location platform developed by Yahoo, called *Yahoo! GeoPlanet*, allows users to look up the unique identifier, called the Where on Earth ID, or WOEID, for almost any named place on the Earth. *Yahoo! GeoPlanet* also allows users to resolve a WOEID they have received from a third party to the place it represents [207]. Due to the increasing availability of smartphones and GPS-enabled devices, the amount of geotagged data content grows exponentially. It paves the way for methodologies and applications of robust social media learning at large scales. Geo-oriented research and applications have successfully grown for more than half of a decade, as can be witnessed by the blossom of location-aware publications.

According to [124] in a 2015 report, location-based services and real-time location applications have become a huge profitable business, with the global location based services (LBS) market expected to grow from USD 11.36 billion in 2015 to USD 54.95 billion by 2020, at a compound annual growth rate of 37.1%. Several factors such as increasing availability of spatial data and analytical tools, growing focus on business intelligence, increasing market competitiveness, navigation, local search, mobile advertisements, location-specific health information, consumer tracking, tourism services, and decreasing cost of solutions are expected to sustain the growth of this market. According to [69] in 2015 market report, almost 3 billion mobile applications are currently in use that rely on location information. In 2013, at the first time ever, more than 1 billion smartphones were shipped worldwide; however, not only smartphones or tablets,

but also other GPS-enabled devices, e.g. tracking devices, digital cameras, and fitness gears demand the need of georeferenced technologies.

1.3 Unveiling the Power of Topics Modeling

Recent years have brought a dramatic increase in the volume of online information. Massive quantities of public text data are readily available across the Web in a wide variety of domains and formations, including news articles, webblogs, forum discussions, social media status updates, reviews of products and services, and scientific publications. According to Google's latest report[1], over 60 trillion individual pages have been indexed. The majority of data are created by individual users via social media. It is seemingly a never-ending feed of information. 3 billion people access to the Internet daily and approximately create 2.5 quintillion bytes of data in which 90% of data generated is unstructured.

The massive amount of data generated on common social media sites are laden with the opinions of individuals regarding to diverse subject ranging from personal to global issues. Topic mining and modeling on social media data can discover and recognize these topics of interest. This fact underscores the relevance of data mining, text analytics, and machine learning techniques in the domain of topic discovery. Hence, advanced programs and models are required to understand exactly what individuals share and to recommend the appropriate results based on individuals' information needs. As a result, this is where we direct our primary effort.

Various methods have been developed to analyze the topics which arise from social media sources. Topic models crawl through a large body of input text to discover latent topics. Topics revealed by the topic model can also be used to feed other analytical tasks such as discovering user interests [104, 116, 206], detecting emerging topics in social media [89, 137, 165, 192], or summarizing a text collection [1, 150]. The topics uncovered can also be used to provide training labels to explore other text sources. Recent advances in topic modeling allow algorithms to be developed that can be used with streaming data from Twitter and another data streams, making this technique an increasingly important analytic tool. Undeniably, topic modeling has become one of the most important techniques in data mining, text mining and machine learning.

The most inspiring contribution of topic modeling is to automatically discover and classify documents in a collection of texts by a number of topics, to represent every document with multiple topics and their corresponding distribution. The topic-based representation generated by using a topic model can solve the problem of semantic confusion. Nevertheless, there are two problems in directly applying a topic model:

- The first problem is the topic distribution within a document itself. It is challenging to strictly assign a document to only one topic, or insufficient to represent

[1] http://www.google.com/insidesearch/howsearchworks/thestory/

it in dimensional representations, a pre-specified number of topics, because some specific and detailed information might be ignored.

- The second problem is that word-based topic representations, a topic is represented by a set of words, are depended on distinctive and semantical differentiation. In other words, the same words that appear in different order contribute to different topics.

Therefore, a topic model needs to take these two problems into account and new techniques are thus demanded to improve the topic interpretations. In this thesis, a conceptual stability based topic model is proposed to enhance the semantic interpretations of topics.

1.4 Summary of Problems Addressed

The two problems of geolocation and topic modeling are exactly what this dissertation addresses. We are primarily concerned with geolocation and topic modeling as means to discover the characteristics and properties of locations and individuals in social media. Certainly, they are many other social media and Internet-related problems that have been researched in the literature, but are not the focus of this thesis.

Over the next sections, we provide a general overview of machine learning methods including supervised and unsupervised techniques which serve as the theory background to solve the focused tasks in this thesis. By concentrating on the key problems, one of the goals of this thesis is to pave the way for development of robust theoretical analysis and application in social media at large scales. We discuss some research contributions, publications and provide a chapter overview.

1.5 Main Research Contributions

During the research of this dissertation, we have made a number of contributions in several fields. Although a number of supervised models for geolocation and unsupervised approaches for topic modeling have been proposed in the last years, there are still several undiscovered scenarios and gaps in the state-of-the-art which need to be investigated. The main goal of this thesis is to provide a cohesive view of the current state-of-the-art technologies, identify unsolved machine learning problems, and propose novel approaches for geolocation and topic modeling. Table 1.1 presents the research contributions that are targeted per chapter. Specifically, the contributions of this dissertation are summarized as follows:

1. **Address a new scenario in geolocation**. We propose a formalization of a matrix factorization based regression model to predict real-time geolocation of

Table 1.1: An overview of the main contributions per chapter in this dissertation.

	Ch.5	Ch.6	Ch.7
Georeferencing real-time location in Twitter streams	•		
Modeling conceptual topics and document clustering		•	•
Discovering hierarchy of topics and spatial distribution			•

Twitter users, an issue which has not been addressed in the literature. The real-time scenario is very important if the location of a Twitter user needed to be known right after posting a tweet.

2. **Propose a new approach for unsupervised topic modeling and document clustering**. We introduce a novel conceptual stability analysis framework to address the challenging question on discovering the appropriate number of topics for a textual source. Our proposed framework has outperformed current state-of-the-art topic modeling approaches.

3. **Develop a general framework for understanding hierarchy of topics and their spatial distribution for streaming data**. The intensive implementation and demonstration of our approach has proved its generalization and practicability. The framework is very flexible in different languages and designed hierarchical topic structure.

Social media learning is an emerging research field where there are certainly different interesting problems that have been studied in the literature which is not in the focus of this thesis. However, equipped with interdisciplinary concepts and theories, new computing scenarios and applications introduced, and several state-of-the-art algorithms proposed, we firmly devote our contributions to both social media and machine learning communities.

1.6 Publications

The contributions of this thesis were published and submitted in several international conferences, challenges and workshops. These publications provide an integral and consistent achievement of the work. Every chapter of the dissertation is based on several works, as enlisted below:

1. Nghia Duong-Trung, Martin Wistuba, Lucas Rego Drumond, Lars Schmidt-Thieme (2015). Geo_ML@MediaEval Placing Task 2015. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop. Wurzen, Germany.

2. Nghia Duong-Trung, Nicolas Schilling, Lucas Rego Drumond, Lars Schmidt-Thieme (2017): An Effective Approach for Geolocation Prediction in Twitter

Streams Using Clustering Based Discretization, European Conference on Data Analysis, in Proceedings of Journal of Archives of Data Science, Series A, ISSN: 2363-9881, Colchester, United Kingdom.

3. Nghia Duong-Trung, Nicolas Schilling, Lucas Rego Drumond, Lars Schmidt-Thieme (2016): Matrix Factorization for Near Real-time Geolocation Prediction, in Proceedings of Lernen Wissen Daten Analysen (LWDA 2016), Potsdam, Germany.

4. Martin Wistuba, Nghia Duong-Trung, Nicolas Schilling, and Lars Schmidt-Thieme (2016). Bank Card Usage Prediction Exploiting Geolocation Information. In Proceedings of the ECML/PKDD Discovery Challenge 2016 on Bank Card Usage Analysis (ECML/PKDD DC 2016). Riva del Garda, Italy.

5. Nghia Duong-Trung, Nicolas Schilling, Lars Schmidt-Thieme (2016). Near Real-time Geolocation Prediction in Twitter Streams via Matrix Factorization Based Regression. In Proceedings of the 25^{th} ACM International Conference on Information and Knowledge Management (CIKM 2016). Indianapolis, United States of America.

1.7 Chapter Overview

The thesis is organized as follows:

- Fist of all, **Chapter 1** serves as a brief introduction of thesis' scope. We present the context of social media learning in general, why it is important and how machine learning can address the most interesting problems. Within this chapter, we give an overview of the remainder of this dissertation and explain how the different chapters are linked together.

- In **Chapter 2**, we formalize the problem description that we cover and solve in this thesis. From the supervised learning perspective, we review geolocalization prediction problem in two research scenarios: user location prediction and real-time location prediction. From the unsupervised learning perspective, we introduce the document clustering and topic modeling problem. The work done in the **publication 4** is a good example of a real world problem that encourages our research in geolocation.

- As we mostly work with textual datasets, proper data cleaning techniques should be carefully considered. In **Chapter 3**, we summarize implemented datasets and present our data-preprocessing procedures.

- **Chapter 4** investigates a simple but effective method to address the user geolocation prediction problem. In this chapter, we develop a clustering based discretization approach which is an effective combination of three well-known

machine learning algorithms: k-means clustering, support vector machines and k-nearest neighbor. Our empirical results indicate that our approach outperforms previous attempts on publicly available datasets, achieving state-of-the-art performance. This chapter is developed upon the work in the **publication 1** and the **publication 2**.

- In **Chapter 5** we introduce a new geolocation scenario that has not existed in previous literature. More specifically, previous research on content-based geolocation in general has addressed the user geolocation prediction problem. Tweets are crawled within a duration of time, e.g. days, weeks. Then, they are concatenated into representative documents to predict users' geolocation. In this work, we develop a novel, generative content-based regression model via matrix factorization approximation to tackle the problem of real-time geolocation prediction. Our method can be accomplished if we leave out the concatenation of tweets. The evaluation of our model shows that our proposed method achieves state-of-the-art performance. This chapter is developed upon the work in the **publication 3** and **publication 5**.

- A novel document clustering approach in text mining is introduced in **Chapter 6**. This chapter investigates the idea of stability analysis, random perturbations, and semantic measurement with which we propose a state-of-the-art topic modeling. We present a challenging question on discovering the appropriate number of topics for a text source and we explain why it has not been successfully addressed in previous literature. Through this, we propose a novel conceptual stability analysis in conjunction with nonnegative matrix factorization to guide the selection of the appropriate number of topics for a specific textual source. This chapter is developed upon the work in the **submission 7** and **submission 8**.

- **Chapter 7** extends the work developed in **Chapter 6** and builds a general framework for topic discovery and spatial distribution in Twitter data. Our motivation is to use a hierarchical topic modeling technique to capture the nature of the content posted on Twitter which varies according to locations. Given a collection of tweets within a region, we aim to address two challenging questions in one useful framework: "how many topics are there?" and "how they are geographically distributed?". The classification and visualization are combined to provide a valuable framework for not only investigating the content and coverage of Twitter usage but also textual sources in general. This chapter is developed upon the work in the **submission 6**.

- Finally, **Chapter 8** summarizes all the proposed approaches. We also discuss an outlook in these research areas and works for the future.

Chapter 2

Technical Basics

Contents

2.1 Overview of Machine Learning

This section presents a critical review of the literature essential to addressing the general problem description and a set of selected methods that are useful to understand the theories and implementation aspects conducted in the thesis. We briefly introduce supervised learning models that support the geolocation task and unsupervised learning techniques to assist the understanding of topic clustering. With provided information, it is not an introduction to the field of machine learning. In order to fully understand the forthcoming concepts, one needs to have acquired a general knowledge of machine learning, which is commonly found in relevant textbooks [20, 62, 63, 64, 117, 125, 134, 138, 156, 163]. Based on the overview of machine learning, we also review several learning algorithms that might help understand problems discussed in next chapters.

2.2 Supervised Learning

Suppose that a data point is a tuple in the format $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^m$ is a vector represented by m features and $y \in \mathbb{R}$ is the associated label. We seek to learn a machine learning model f that maps $\mathbf{x}$ to y:

$$f : \mathbb{R}^m \rightarrow \mathbb{R}. \tag{2.1}$$

In other words, our task is to find a mapping $f(\mathbf{x})$ such that $y = f(\mathbf{x})$. After the model f is trained using the training set, we are given an unlabeled set or the *test* set, in which observations are in the format $(\mathbf{v}, \hat{y})$ where $\hat{y}$ is the prediction made by f given $\mathbf{v}$. We compute $\hat{y}_j = f(\mathbf{x}_j)$ which is the predicted label of the unlabeled observation. We use the hat symbol to denote an estimate.

Supervised learning can be divided into *classification* and *regression* depending on the value of labels. When a label's value is discrete, the problem is called classification. Here the goal is to learn a mapping from inputs $\mathbf{x}$ to outputs y, where $y \in \{1, \ldots, C\}$, with C being the number of classes. If $C = 2$, this is called binary classification. For convenience, we usually denote $y \in \{0, 1\}$ in this case. If $C > 2$, this is called multiclass classification. If the class labels are not mutually exclusive, it is called multi-label classification. Otherwise, when the label's value is continuous, it is called regression.

In this thesis, we introduce a classification method for the user geolocation problem and a regression method for the real-time geolocation problem. For the task of predicting a tweet's geolocation, the labels are continuous values representing either latitude or longitude. The model starts with a training set, e.g. a collection of tweets, where both features and labels are known. A supervised learning algorithm is trained on the training set in a process known as *induction*. where a model f is learned. Then, the model f is evaluated on a test set in which each unlabeled observation is assigned a predicted label. This process is called the *deduction*.

2.3 Supervised Learning Evaluation

Supervised learning algorithms often employ a training-validation-test framework in which a training set is used to train a model, a validation set is used to select the best parameters and hyperparameters, and a test set is used to evaluate the model. The performance of supervised learning algorithms are measured by how accurate they are in predicting the correct labels of the test set. In practice, the labels of a test set are unknown. However, we can create a test set and called it as a validation set to pick the model of right complexity. Hence, the training set is divided into two parts, one used for training the model and other used for model validation. We then fit the model on the training set, and evaluate its performance on the validation set, and search for the best parameters and hyperparameters. After the labels of the validation set are predicted using the learning model, the predicted labels are compared with the ground-truth. This measures how well the trained model is generalized to predict class values. If we have a separate test set, we can evaluate performance on this, in order to estimate the accuracy of the model.

We often use about 80% of the data for the training set, and 20% for the validation set. But with this scenario, some *good* instances may be split into the validation set instead of the training set. A simple but popular solution to this scenario is to use a cross validation technique. We split the training set into $K \in \mathbb{N}^+$ folds; then, for each fold $k \in \{1, \ldots, K\}$, we train the model on all the folds but the k-th, and test it on the k-th, in a round-robin fashion. The average performance of the model over k rounds measures the generalization accuracy of the model. This robust technique is known as *k-folds cross validation*. It is common to use $K = 5$, e.g. 5-folds cross validation. If we set $K = n$ many instances, then we get a method called *leave-one out cross validation*.

different evaluation techniques can be applied to compare the ground-truth labels and the predicted labels, depending on the type of supervised learning algorithm. In classification, the class values are discrete which allow us to use the accuracy to evaluate the model. Basically, the accuracy is the fraction of labels that are predicted correctly over the total labels. In regression, it is unreasonable to assume that the label can be predicted precisely because their labels are real values. Therefore, we check if the predictions are highly correlated with the ground-truth.

2.4 Unsupervised Learning

We now discuss unsupervised learning, where the goal is to discover interesting structures in the data that sometimes is considered as knowledge discovery. Unlike supervised learning, we are not told what the desired output is for each input. Generally speaking, unsupervised learning is the division of instances into groups of similar objects. In this thesis, we focus on feature-based clustering where the input to the model is an $n \times m$ feature matrix.