Twitter Stream Mining: analysis of basic techniques

Christina Pardalidou
Department of Informatics
Aristotle University of Thessaloniki
{chripard@csd.auth.gr}

Alex Tsilingiris
Department of Informatics
Aristotle University of Thessaloniki
{alextsil@csd.auth.gr}

Sotiris Karapostolakis
Department of Informatics
Aristotle University of Thessaloniki
{skarapost@csd.auth.gr}

ABSTRACT

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. Detecting topics in Twitter streams has been gaining an increasing amount of attention. It can be of great support for communities struck by natural disasters, and could assist companies and political parties understand users' opinions and needs. Sentiment analysis of twitter data is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. Another interesting topic regarding twitter posts is the discovery of their geolocation. The limited availability of self-identified geolocation information makes it challenging to place current and trending topics on the globe. In this paper, we present various techniques concerning all of the topics mentioned above.

Keywords

Twitter, Event Detection, Sentiment Analysis, Geolocation

1. INTRODUCTION

Nowadays, the World Wide Web has transformed from a large, static library that people only browse into a vast and dynamic information resource. Relying on this, social networks is a very popular and powerful tool for expressing opinions, broadcasting news, and simply communicating with friends. People using them for commenting on significant events in real time, with several hundred micro-blogs posted each second.

The most popular micro-blogging service is Twitter. The popularity of Twitter stems from its availability on a number of different electronic devices (web and cell phones. There is a prevalence of a subculture in Twitter that encourages users to acquire a large friend pool, as well as send tweets on a wide variety of subjects, typically several times a day.

Monitoring and analysing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets. Tweets can be seen as a dynamic source of information enabling individuals, corporations, and government organizations to stay informed of "what is happening now." For instance, people would be interested in getting advice, opinions, facts, or updates on news or events[30]. Companies are increasingly using Twitter to advertise and recommend products, brands, and services; to build and maintain reputations; to analyse users' sentiment regarding their products (or those of their competitors); to respond to customers' complaints; and to improve decision making and business intelligence. Twitter has also emerged as a fast communication channel for gathering and spreading breaking news, for predicting election results, and for sharing political events and conversations. It has also become an important analytical tool for crime prediction and monitoring terrorist activities.

Twitter promotes an attractive style stating breaking news, as there is very little lag[32] between the time that an event happens or is first reported in the news media and the time at which it is the subject of a posting on Twitter. Twitter can be characterized as an endless database, which collects millions of real-time short text messages every second. Tweets also have a mechanism by which the user can link to other objects on the web such as articles, images or videos which is typically used to link tweets to related material on the Internet. Thereafter, the first result is that the size of information is multiplied and the variety of references is bigger, as well. These messages are not only just data, but they can be manipulated efficiently. One well-timed subject of research is to use those messages for event detection. In other words, the tweets is a source of inventing which topics are more seasonable. Event detection has also instant impact on the world, through the quick transmission of the news and necessary briefing in some cases.

With the passage of time and the effect of more and more users the topic acquire much popularity. Through this phenomenon we can form a general summarization of the event. This process is called event summarization. The massive crowd keeps close pace with the development of trending topics and provide the timely updated information. Twitter has shown its powerful ability in information delivery in many events, like the wildfires in San Diego and the earthquake in Japan[13]. In response to searches for ongoing events, today's major search engines simply find tweets that match the query terms, and present the most recent ones. This approach has the advantage of leveraging existing query matching technologies, and for simple one-shot events such as earthquakes it works well. However, for events that have "structure" or are long-running, and where users are likely to want a summary of all occurrences so far, this approach is often unsatisfactory.

Event detection is a growing domain of research. Many different species of algorithms have been detected regarding this sector. A common approach are techniques that are based on text categorization. In addition, there are some methods that reclaim the display frequency of each term.

The computational treatment of sentiment has recently attracted a great deal of attention, in part because of its potential applications. One of the main reasons for sentiment analysis is the aforementioned increase of user-generated content on the Web which has resulted in a wealth of information that is potentially of vital importance to institutions and companies. Typically, document-based sentiment analysis processes operate at a particular level, i.e. at the word or sentence level, for extracting a document's sentiment. In machine learning, the most popular approach for sentiment analysis, the selection of appropriate features for representing a document is crucial. In sentiment identification at the word level different types of features have been introduced, which are either sentiment-based (e.g. words which express a specific sentiment), syntactic-based (e.g. part-of-speech and n-grams)[8], or semantic-based (e.g. semantic word vector spaces which capture the meaning of each word).

Document-level polarity classification is not a special case of text categorization with sentiment -rather than topic-based categories. Hence, standard machine learning classification techniques, such as support vector machines (SVMs) [29], can be applied to the entire documents themselves. Nevertheless, some researches presented a technique that it is easy to improve the accuracy, by integrating sentence-level subjectivity detection with document-level sentiment polarity.

As we know, in the machine learning approach, each classifier is trained using a collection of representative data. In contrast, the semantic-orientation approach does not require prior training; instead, it measures a word containing positive or negative sentiment. Each approach has its own benefits and drawbacks. For example, the machine learning approach tends to be more accurate, but the semantic-orientation approach has better generality[11]. Recently, a new lexicon-enhanced method was accrued to generate a set of sentiment words based on a sentiment lexicon as a new feature dimension. It combines these sentiment features with content-free and content-specific features used in the existing machine-learning approach. In the evaluation stage, they showed that adding the new set of sentiment features can increase sentiment-classification performance.

The Internet and other communication technologies play a potentially disruptive role on the constraints imposed on social networks. These technologies reduce the overhead and cost for being introduced to new people regardless of geography, and help us stay in touch with those we know. Some have even gone so far as to call this "the end of geography", where the process of relationship formation becomes disentangled from distance altogether.

However, geography still plays an important role. The reason is because of the strong relationship between event detection and geographical location each user belongs. Twitter is a social networking website, which means that users need not be viewed in isolation, but instead can be viewed as part of a large network of other users, user groups, and user cliques. Moreover, users have some meta-data information, such as description, source location, friends, which means that the social network structure in Twitter can aid in finding users that are most likely to tweet about news belonging to a particular geographic location or region.

The rise of micro-blogging services spurred various applications to mine the data coming from those services[26]. Many such applications could benefit from information about the location of users, but unfortunately location information is currently very sparse. The main problem is that less than 1 per cent of tweets are geo-tagged and information available from the location field in users' profiles is unreliable at best. The benefits of mining those data promises new personalized information services, including local news summarized from tweets of nearby Twitter users, the targeting of regional advertisements, spreading business information to local customers, and novel location-based applications (e.g., Twitter-based earthquake detection, which can be faster than through traditional official channels).

There is a great number of geoinference using social networks One direction has produced approaches that claim to accurately locate the majority of posts within tens of kilometres of their true locations[17]. Another method predicts the location of an individual from a sparse set of located users with peformance that exceeds IP-based geolocation. On the side, there is also a technique that predicts locations of Twitter users at different granularities, such as city, state, or time zone, using the content of their tweets and their tweeting behaviour.

Getting started the first section of this survey is the presentation of some techniques that aim to event detection. The survey detects both algorithms that are based on text categorization and frequency display methods. The second chapter deals with techniques of sentiment analysis. We give more weight on techniques that use machine learning. The last chapter unfolds methods for location identification.

2. EVENT DETECTION

Given a series of twitter posts, the goal of event detection is to extract a particular event by analysing the text or hashtag of a tweet. The process of event detection is not a lenient task as tweets stream in huge volumes and the level of noise is kept high [30]. On the other hand, the huge volume of the stream allows the use of streaming algorithms, thus making event detection an accomplishable task [30].

The conventional approach for this problem is to represent the documents as term frequency vectors [30]. When a new document arrives, it is compared to all previous ones, and if its similarity to a specific document called "centroid" is below a threshold, the new document is registered as a new event [30]. Unfortunately, this simple approach doesn't perform well as the dimensionality of the data expands.

Twitter's users usually tweet about their personal life, personal opinions, taste in music etc., which makes the majority of tweets unsuitable for our purpose. Although, when a noteworthy event occurs, a lot of users tweet about it. The goal of event detection is to detect these events in an automatic fashion, keeping the noise (tweets that are not events) as low as possible [30].

2.1 Strategies for event detection

As stated before, tweets stream in high rates and capture the pulse of the moment, therefore occupying us to process them right away [32]. This is accomplished by using algorithms that are online in nature, meaning they can on inputs that are not known beforehand[32].

Tweets are inherently noisy since most of them concern small user groups. In order to improve the results of event detection we first have to improve the input by extracting only useful information from noise [32]. Even then, information that concerns us is still limited by the upper character limit set for each tweet, currently at 140 characters. Bad use of language, spelling errors and special puns also negatively affect the outcome of event detection.

Twitter is an evolving platform where users come and go, and become friends and followers with each other. When designing event detection algorithms, one should keep in mind the above aspects and make the algorithms adapt well to new knowledge, otherwise watch them fail as time passes [32].

The difference between newswire and Tweeter is that, in Tweeter, the identities of the news "reporters" in not known in advance and the news are not posted based on a schedule, but rather appear randomly among other tweets [32].

2.1.1 *Identifying users*

An approach to event detection is the identification of users that usually tweet about news [32]. A technique to do this is to manually identify these users by looking for the most common set of followers among them [32]. Another technique to do this is to assume that a user with a high amount of followers represents an influential event source into a social community [6]. As an example, we present figure 1 , which features Barack Obama as the most influential node in this group [6].

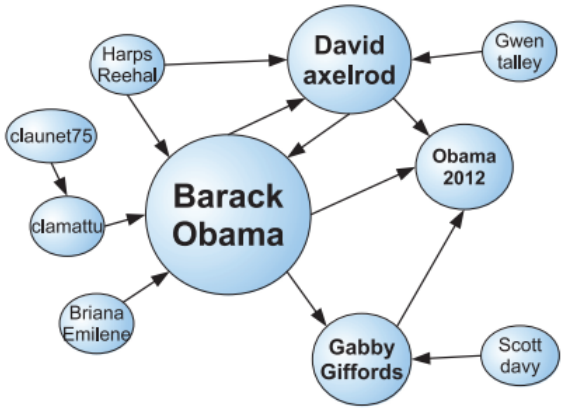

Figure 1: Graph showing user importance on Twitter, based on the number of their followers. [6]

The added benefit of this method is that, not only we identify the users of interest but, we get an evaluation of a user's influence, meaning that an influential user can affect other users' tweet streams [6]. This technique cascades as the follow-like relationships spread, making it comparable to the PageRank algorithm [6]. An interesting fact to support the above claims is that studies have shown that 10% of Twitter's users are responsible for 90% of the tweets [32].

2.2 Term and Topic Extraction

Although manual techniques exists, we will solely focus on automatic extraction techniques as they fit best on the Twitter paradigm. Filtering could be applied on all tweets, expect those posted by influential users that usually post

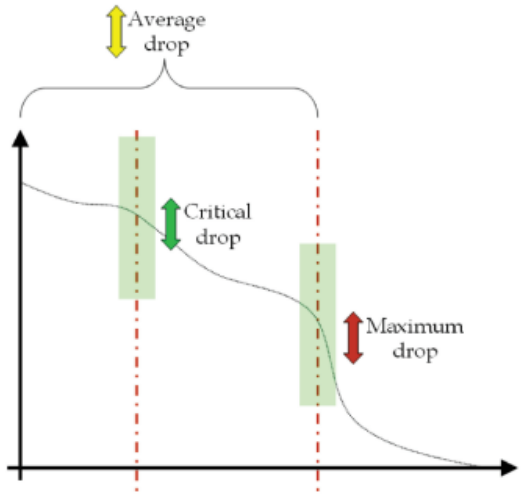

Figure 2: The critical drop cut-off: the maximum drop is the highest variation in the ordered list of weights (red mark). The average drop (between consecutive entities) is the average difference between those items that are ranked before the identified maximum drop point (yellow mark). The first drop which is higher than the computed average drop is called the critical drop (green mark). Figure also taken from [6].

news, leaving us with a guaranteed source of news. The goal of the extraction is to find a way to distinguish tweets that are clearly not news, and not completely eliminate the noise, which is an enormous task to take on [32]. For the above task, [32] used a naive Bayes classifier that was trained using tweets that were marked as news or not news.

2.2.1 *Setting a threshold for term extraction*

An approach to selecting emerging terms relies on a user-specified threshold parameter, as described by [6]. It is assumed that, given two keywords that are likely to be news, they can be considered as such based on user evaluation. A threshold value called critical drop is introduced, which allows the user to decide when a term is news or not.

In order for the above technique to work, a proper drop value has to be set. This a be a daunting task for a user or a machine. The difficulty lies in the fact that it's difficult to quantify a threshold given an abstract context that is, on top of other things, unstable. An automatic way of doing this, as described by [6], is shown below: The main idea is to preserve terms with high frequencies but only consider those whose frequency is higher than the average frequencies of most discussed terms. A model is introduced that works as follows:

- Terms are ranked in descending order based on their frequencies
- The maximum drop value between consecutive entries is calculated and the drop point is identified.
- The average drop between consecutive entities for all those keywords that are ranked before the identified maximum drop point is calculated.

- The first drop which is higher than the computed average drop is identified as the critical drop.

At the end of this procedure, the terms ranked before the critical drop are defined as news on the specified time interval.

2.2.2 Clustering

The goal of clustering techniques is to group news tweets such that each group contains tweets about a specific news topic. This problem is similar to clustering documents with the main difference that the identity of news topics is not known beforehand [32]. This problem cannot be approached by static training sets, as news evolve over time in a random fashion. A clustering approach, presented by [32], goes as follows: Once a tweet arrives, it is added to a news cluster and remains there forever. This techniques allows for fast processing since tweets arrive at huge rates. This makes the algorithm online. Specifically, an algorithm called leader-follower clustering has been modified and used, which allowed for clustering both in content and time. Along with the clusters, a weighted list of keywords of each tweet is kept. Time is also an important aspect in this procedure: Timestamps are kept and can render a cluster inactive if its centroid is older than a specified threshold. In this case the cluster is locked and new tweets cannot be added.

Noise can greatly affect clustering techniques. Care should be taken when creating new clusters to ensure they are of good quality. A good way to ensure high quality is to only allow the creation of clusters by influential users with many followers [32]. All tweets from other users should be able to join a cluster but not create one. Unfortunately, this approach might prevent news registration for cases where simple users report an incident first. A way to combat this situation effectively is to allow cluster creation by users, but only activate a cluster after a certain amount of tweets has been added to it, thus making the topic news-worthy [32].

2.2.3 A different approach on Clustering

As [6] states, many clustering and classification strategies cannot by applies to this task due to the fact that they tend to ignore relationships among documents(tweets) related to a news event. [6] proposes a metaphor where each term is seen as a living organism: A term's vitality status increases as users use it, and slowly degrades as references drop over time. The term's vitality is weighted based on the source of the tweet. As stated before, highly influential users are more likely to post high-importance news, hence allowing us to favour terms found in their tweets.

3. SENTIMENT ANALYSIS

Sentiment analysis (SA) can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP)[21]. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. Twitter Sentiment Analysis (TSA) tackles the problem of analyzing the messages posted on Twitter in terms of the sentiments they express. Twitter is a novel domain for SA and very challenging. One of the main challenges is the length limitation, according to which tweets can be up to 140 characters. In addition, the short length and the informal type of the medium have caused the emergence of textual informalities that are extensively encountered in Twitter. Thus, methods proposed for TSA should take into account these unique characteristics. The majority of TSA methods use a method from the field of machine learning, known as classifier. Figure 3 shows the most typical TSA process.

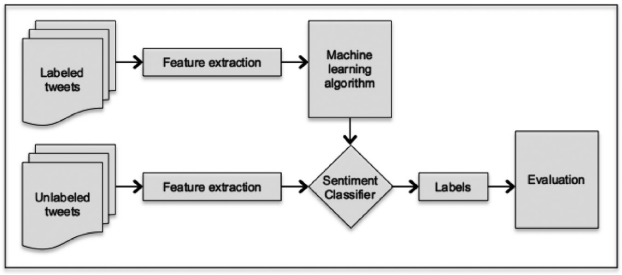

Figure 3: Typical process for sentiment classification. [15].

In the literature, SA has been applied at three different levels: *document*, *sentence*, and *entity levels*. SA at the document level aims to identify the sentiment polarity expressed in the whole document. The sentence level SA aims to classify each sentence as positive or negative, whereas entity-level SA detects the sentiment polarity of a specific entity/target of a particular object. Due to the length limitation, the majority of tweets contains a single sentence. Therefore, for the task of TSA there is no fundamental difference between document and sentence level. In case of tweets, SA can be applied on two levels: *message/sentence* and *entity levels*.

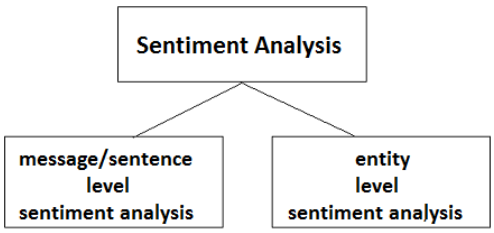

Figure 4: Levels of sentiment analysis related to tweets.

Four different classes can be identified in the literature of TSA:

- Machine Learning
- Lexicon-Based
- Hybrid (Machine Learning and Lexicon-Based)
- Graph-Based

3.1 Machine Learning

The majority of the proposed methods that deal with TSA employs a classifier from the field of machine learning that is trained on various features of tweets. Some of the most applied classifiers are the Naive Bayes (NB), Maximum Entropy (MaxEnt), Support Vector Machines (SVM), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Random Forest (RF), and Conditional Random Field (CRF).

[35] proposes learning *sentiment specific word embedding* (SSWE) for sentiment analysis. It develops *neural* networks and maps each *ngram* to the sentiment polarity of sentence. This allows it to focus on learning the meaning of word, namely word embedding, from massive distant-supervised tweets by doing this from scratch. The effectiveness of SSWE has been implicitly evaluated by using it as features in sentiment classification on the benchmark dataset in *SemEval 2013* [28], and explicitly verified by measuring word similarity in the embedding space for sentiment lexicons [35].

A deep learning approach that also uses neural networks is presented in [33]. This approach focuses on sentiment analysis on both message and phrase levels. The resulting model of this paper demonstrates state-of-the-art performance on both levels.

One of the most used classifiers for addressing TSA is the SVM classifier. Bakliwal et al. 2012 [2] employed an SVM classifier trained on 11 features to address TSA. They employed different pre-processing techniques one by one in order to measure their effectiveness. *Spelling correction*, *stemming*, and *stop-words removal* managed to increase the accuracy of the classifier. Two different datasets were used to evaluate their approach, the Stanford dataset [16] and the Mejaj [4]. The best results were achieved by the combination of NLP- and Twitter-specific features.

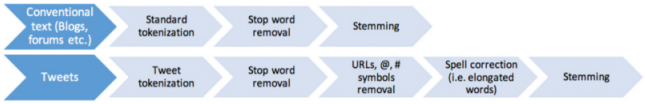

Figure 5: Typical scenarios of preprocessing on standard text and tweets [15].

3.2 Lexicon-Based

Lexicon-based methods leverage lists of words annotated by polarity or polarity score to determine the overall opinion score of a given text. The main advantage of these methods is that they do not require training data. Lexicon-based approaches have been extensively applied on conventional text such as blogs, forums, and product reviews. However, they are less explored in TSA compared to machine-learning methods. The main reason is the uniqueness of the text on Twitter that not only contains a large number of textual peculiarities and colloquial expressions such as *yolo* and *gr8* but also has a dynamic nature with new expressions and hashtags emerging from time to time.

Ortega et al. 2013 [5] proposed a three-step technique for TSA. Pre-processing was performed in the first step and polarity detection in the second step. In the last step, they performed rule-based classification. Polarity detection and rule-based classification were based on WordNet and SentiWordNet. Their approach managed to achieve good results when evaluated on the SemEval-2013 dataset [28]. However,the authors did not compare their method with existing techniques to prove its effectiveness. [11] introduced some new sentiment features. The results showed that adding these newly created features, which are often used in the existing semantic-orientation approach, and the content-free and content-specific features that come from the existing machine-learning approach can improve sentiment - classification performance significantly.

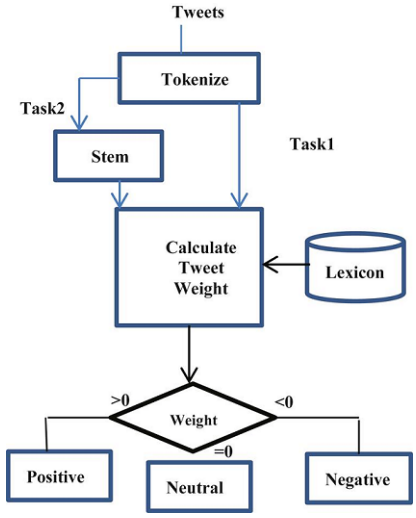

Figure 6: An example of the architecture of a framework of lexicon-based sentiment analysis.

Another approach using a newly proposed methodology, that is lexicon-based, for capturing people's emotions in real time is presented in [7]. This automated technique provides the means for an automated real-time analysis. The evaluation of the experimental results proved that the proposed method managed to capture efficiently the emotions expressed in tweets as well as their intensity.

3.3 Hybrid

A number of researchers combined machine-learning and lexicon-based approaches.

An interesting hybrid method was presented by Ghiassi et al. 2013 [14], who combined dynamic artificial neural network with n-gram. Emoticons and tweets that contained the word love or hate or their synonyms were used as features to build the two classifiers: SVM and a Dynamic Architecture for Artificial Neural Networks (DAN2). The proposed approach was tested on a collection of tweets crawled using the subject *Justin Bieber*. The results showed that DAN2 managed to outperform SVM.

Khuc et al. 2012 [22] also combined a lexicon-based approach with a classifier to improve TSA accuracy. They considered the MapReduce framework to create a co-occurrence matrix based on bigram phrases. The cosine similarity between words is then computed and the edges with low cosine score are removed. Then, they combined the score generated using a simple lexicon-based approach with a classifier. For the machine learning algorithm, they used the Online LR approach. Their experiments showed that the hybrid approach outperformed the simple lexicon-based classifier that was only based on words/phrases that indicated sentiment.

Another older interesting study was presented by Zhang et al. 2011 [37], who proposed a hybrid method to address entity-based TSA. For each of the entities Obama, Harry Potter, Tangled, iPad, and Packers they computed a sentiment score based on their proximity to words from a sentiment lexicon. They proposed a rule-based algorithm that also considered *comparative judgments*, *negation*, and *expressions* that were likely to change the orientation of a phrase. To collect more annotated data and enhance the recall of the proposed method, they identified additional subjective terms using *Chi-square*. The SVM classifier was then applied for sentiment polarity detection.

3.4 Graph-Based

Although machine-learning methods achieve a great performance on TSA, they require a large number of annotated data. Label propagation is a method that can reduce the demand of the annotated data. For this reason, a number of researchers utilized the Twitter social graph under the assumption that people influence one another. Label propagation is a *semi-supervised* method in which labels are distributed to nodes using the connection graphs.

Some of the first to apply a label propagation method for TSA were Speriosu et al. 2011 [34]. Their proposed method leveraged the Twitter follower graph under the assumption that people influence one another. Users, tweets, unigrams, bigrams, hashtags, and emoticons were used as nodes for the construction of the graph. The proposed label propagation method outperformed a lexicon-based approach and a MaxEnt classifier.

Cui et al. 2011 [10] tackled TSA with a label propagation method based on analysis of emotion tokens. At first, the emotion tokens from tweets were extracted. A graph propagation method was then used to assign polarities to the tokens. In the final step, they analyzed and classified the emotion tokens. The emotion tokens included *emoticons*, *repeating punctuations*, and *repeating letters*. Their approach managed to perform well in analyzing sentiment of messages written in any natural language.

Instead of using emotion tokens, Wang et al. 2011 [36] proposed a graph-based model that leveraged co-occurrence of hashtags to classify the sentiment of certain hashtags. They proposed different algorithms that were compared to an SVM voting. The SVM was trained with several features, including unigrams, punctuation, and emoticons. The Loopy Belief Propagation algorithm managed to achieve the best performance in terms of accuracy compared to the other tested methods.

4. GEOLOCATION PREDICTION

Social media produces a continuously-updated stream of data which has proven useful for predicting group behaviors and modeling populations. Associating data with the particular geolocation from which it originated creates a powerful tool. Geo-tagged tweets would be a significant reinforcement to a variety of applications. Therefore[17], recent work has focused on geoinference for predicting the locations of posts. In this section, we are going to epitomize some of the most basic techniques for Geolocation prediction.

4.1 Prediction at Different Granularities

A number of different kind of techniques have been recently proposed. [26] propose a model whom objective is

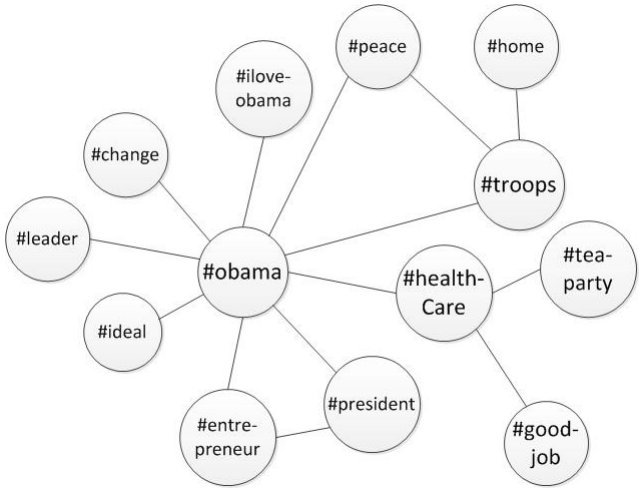

Figure 7: An example of a Hashtag Graph Model [36] .

the prediction of users' location at different granularities. More specifically at city-level. The dataset for this model is consisted of 1,524,522 tweets from 9551 users. The tools for collecting the data are the Google's geocoding API, Twitter's rest API and Twitter's streaming API. Below follows the severance of the techniques to some basic categories.

4.1.1 Content-based Statistical Classifiers

In this part, three different models are trained. We denote as S, the set of all users' tweets. The first one uses all the words contained in S. The second one uses all hashtags that are included in S. The last one, uses all the place names that S contains. The first step is the feature selection. This process is different at each case. For the words the algorithm keeps only the nouns. For the hashtags, the algorithm keeps only the words that start with #. The place names approach keeps only the terms that appear in the tweets and match names of US cities and states from the USGS gazetteer. Next step is the identification of terms that are particularly representative for a location. The last step is the training and classification phase. The learning algorithm that is selected is Naive Bayes Multimonial from Weka.

4.1.2 Content-based Heuristics Classifiers

- Local Place Heuristic Classifier: The heuristic is that a user would report his or her home city or state in tweets more often than any other cities or states.
- Visit Place Heuristic Classifier: The heuristic is that a user would visit places in his home location more often than places in other locations.

4.1.3 Ensemble of Location Classifiers

The main idea in this part is a dynamically weighted ensemble of statistical and heuristic classifiers. Here we have the introduction of the Classification Strength metric. According to this, let T indicate the set of terms from user's tweets that would be considered for classification using a specific classifier. For statistical classifiers, the matching location distribution is the set of locations in our trained

model containing terms from T. For the local-place classifier, this distribution contains locations from our dataset that match content in the user's tweets. For the visit-history classifier, this distribution includes locations from the user's visit history that show up in our dataset. The Classification Strength for a user is the inverse of the number of matching locations in the matching location distribution. In this approach, the Classification Strength of a classifier for a particular instance is used as the weight of that classifier in the ensemble (for classifying that instance) and the location with the highest rank by weighted linear combination is returned as the result.

For the evaluation of the classifiers, the evaluating technique is the Recall(R). The best algorithm was the Place Name Statistical Classifier. The high recall of the place name-based classifier may be explained by the fact that many users send tweets containing names of places (cities and states in our system), and those place names tend to have bias towards users' home cities.

4.2 Network-based Geoinference

The interest in creating new algorithms for geoinference prediction has shown a great increase. A lot of new algorithms that promise great results have been detected[20]. However, the conditions under the conclusions that was assembled are controversial. The objective of this subsection is the assessment of the state-of-the-art techniques. The three evaluating criteria are the accuracy of each algorithm in real-world conditions, (2) the impact of how ground truth is determined for training, and (3) the relationship between performance and time as models infer the locations of posts created after the model was trained. Below is the description of the 9 methods:

4.2.1 Techniques

[1] presented a model using Facebook. In this model this algorithm demands from the users to dispose highly-precised locations. The main idea of this method is the building of a probabilistic model representing the likelihood of observing a relationship between two users given the geographic distance between them. They achieved a complexity of $O(|V|k^2)$ where V is the number of vertices in the graph and k is the average number of vertex neighbors with known locations.

[27] proposed a method citing the difficulties adapting from Facebook to Twitter. The technique is a tree regression model that predicts the actual distance between the users. However, as the size of the network grows, computing the regression model on all pairs of located user dominates the run-time, making the algorithm unpractipable in real-world applications.

[23] proposes several extensions to the [1] model based on strategies for weighting which of a user's friends are likely to be most predictive of their location. This method uses the friendship definition of [18] which is that a user a is friend with b if a has mentioned b in at least two posts. Given a user, their friends are weighted according to a social tightness coefficient, which is computed as the cosine similarity of the two user's friends.

[25] proposes to infer geolocation by taking into account the influence of both users and of locations, capturing the intuition that some users are more informative for predicting the locations of the neighbors. The best-performing of their approaches first assigns users to random locations and then iteratively updates the locations of users from their neighbors and mentioned location names, refining the parameters in the update by measuring the prediction error of users with already-known locations.

[24] recognize that a user may relocate over the course of their social media history and therefore may have more than one home location from which they posts. To uncover all of a user's home locations, relationships between users and locations are modeled with a supervised extension to Latent Dirichlet Allocation[3].

[31] approach geolocation inference as a classification task. a SVM classifier was trained with three core features: (1) the cities of a user's friends, relative to the city's respective number of Twitter users, (2) the number of closed triads in a user's social network residing in the same city, and (3) the number of reciprocal following relationships a user has per city.

[12] propose one of the simplest approaches where given a user, their location is inferred by taking the most-frequently seen location among their social network.

[19] extends the idea of location inference as label propagation by interpreting location labels spatially. Locations are inferred using an iterative, multi-pass procedure. Like [23], during one pass, the location of each user is assigned as the geometeric median of the locations of all the user's neighbors, with groundtruth data providing the locations for the first pass.

[9] extend the method of [19] to take into account edge weights in the social network and to limit the propagation of noisy locations.

4.2.2 Evaluation Phase

The evaluation criteria by which methods should be measured are three: (1) accurate inference of a post's location, (2) accurate inference of all the posts generated by a user, and (3) maximizing the number of posts whose location is able to be inferred. The metrics that are used for the evaluation of the techniques are three, too. Specifically, AUC, Post Coverage and Median-Max.

The first (cross-validation) experiment is an evaluation of all nine methods and the baseline in a common setting, on identical data, using consistent sources of ground truth, and measuring performance with the proposed evaluation metrics.

The second (Self-reported Locations) experiment tests the performance impact of gazetteer choice using four, and, large gazetteers: Geonames, GeoLite, DBpedia and Google.

The last (Temporal Decay) experiment evaluates how each geoinference method changes in accuracy and coverage when predicting locations for novel Twitter data created at increasingly - distant points in time after the last date seen in the training data. Further, this experiment examines the degree to which cross-validation performance is predictive of future performance by the models.

4.3 Prediction with Social and Spatial Proximity

The objective of this subsection is the introduction of an algorithm that predicts the location of an individual from a sparse set of located users with peformance that exceeds IP-based geolocation[1].

It is observed that the likelihood of friendship with a person is decreasing with distance. This should not be surpris-

Method	Geonames			DBpedia			GeoLite			Google		
	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.
Dav11	0.725	28.3km	0.014	0.741	22.7km	0.003	0.690	33.1km	0.019	0.712	20.2km	0.028
Li12a	0.108	8028.0km	0.788	0.094	9375.4km	0.788	0.110	7971.4km	0.788	0.545	8112.0km	0.788
Li12b	0.386	4312.0km	0.617	0.164	7378.3km	0.576	0.221	6720.7km	0.598	0.315	2069.8km	0.261
Rout13	0.217	6520.8km	0.788	0.234	6120.1km	0.788	0.234	6120.1km	0.788	0.231	6167.8km	0.158
McG13	0.492	293.0km	0.078	0.419	344.4km	0.010	0.512	249.4km	0.144	0.520	199.7km	0.133
Kong14	0.447	418.8km	0.534	0.337	806.9km	0.280	0.469	329.8km	0.548	0.483	312.8km	0.538
Back10	0.503	278.5km	0.026	0.428	338.0km	0.024	0.520	246.3km	0.144	0.527	193.7km	0.133
Jurg13	0.477	333.5km	0.623	0.331	948.1km	0.452	0.490	295.7km	0.629	0.512	256.9km	0.624
Comp14	0.509	250.6km	0.556	0.338	916.3km	0.462	0.517	240.8km	0.573	0.543	171.4km	0.566
Rand. N.	0.421	509.7km	0.623	0.310	1215.0km	0.452	0.437	399.7km	0.629	0.455	381.7km	0.624

Figure 8: Self-reported locations identification when varying the gazetteer[20].

Method	GPS			Geonames		
	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.
Dav11	0.633	**57.2km**	0.131	**0.725**	**28.3km**	0.014
Li12a	0.123	7738.9km	**0.788**	0.108	8028.0km	**0.788**
Li12b	0.386	4312.0km	0.617	0.217	6711.3km	0.591
Rout13	0.225	6942.7km	**0.788**	0.217	6520.8km	**0.788**
McG13	0.617	74.6km	0.223	0.492	293.0km	0.078
Kong14	0.596	127.4lkm	0.609	0.447	418.8km	0.534
Back10	0.622	71.9km	0.223	0.503	278.5km	0.026
Jurg13	0.620	102.1km	0.657	0.477	333.5km	0.623
Comp14	**0.656**	61.2km	0.606	0.509	250.6km	0.556
Rand. N.	0.559	182.5km	0.657	0.421	509.7km	0.623

Figure 9: Cross-validation results of the training data, using different sources of ground truth[20].

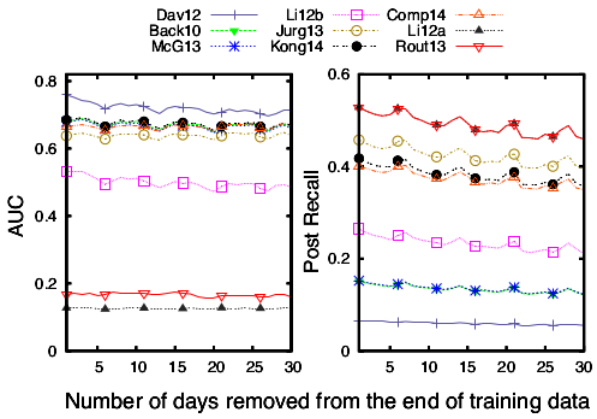

Figure 10: Performance on each day of the test data[20].

ing given that we are less likely to meet people who live further away. A less obvious relationship is that the total number of friends also tends to decrease as distance increases. This means that the probability of knowing someone d miles away is decreasing faster than the total number of people d miles away is increasing.

The model that is proposed finds that using a maximum likelihood approach with the simplifying assumption that the user will be either colocated or in close proximity to one of her friends, it is able to predict the physical location of 69.1 of the users with 16 or more located friends to within 25 miles, compared to only 57.2 using IP-based methods. The next step is to try with even more social data. The content of this data is how often users interact with each other and see each other's content. Using this data generates slight improvement in geolocation, which implies that users who are physically close to each other may tend to interact more often on Facebook. The last step is to iterate the maximum-likelihood algorithm. The reason is to increase the accuracy of prediction of many individuals at once. This showed significant improvement in the accuracy.

5. CONCLUSION

Recent years have witnessed an increasing research interest in analyzing tweets. This interest is a result of the large amount of messages that are posted everyday in Twitter and that contain valuable information for a number of different topics. Event detection aims at finding real-world occurrences that unfold over space and time. As a fast-growing microblogging and online social networking service, Twitter provides unprecedentedly valuable user-generated content that can be transformed into actionable and situational knowledge. A considerable effort is still required to achieve efficient and reliable event detection systems. Some of them are the designing of better feature extraction and query generation techniques, more accurate filtering and detection algorithms, improved techniques to combine and analyze information from multiple sources (social and traditional media) and multiple languages, and enhanced summarization and visualization approaches. As for sentiment extraction, methods from the machine-learning field are applied for TSA on a more frequent rate compared to the rest of the approaches, with the SVM and NB classifiers being the most prevalent. Unigram-based SVM is usually considered as a baseline to which the proposed methods are compared. Moreover, lexicons are utilized in a large set of proposed methods to support detecting words that indicate sentiment. SentiWordNet and MPQA are the most used lexicons that are usually extended with words that are used in Twitter. Continuing, location information on Twitter is available through geotagged tweets indicating the current location of the user at the time of tweeting. However, geotagging is not a default setting and thus rarely used. Accurate user geolocation is a key driver for location-specific services such as localised search, and has been the target of research across different disciplines. One of the most reliable and straightforward approaches to geolocation prediction is IP-based methods, but in many contexts, it is not possible to access the IP of the device used to post content, or the IP is relatively uninformative (as is the case with, e.g. mobile devices). A new-entry algorithm finds the location of a user at different granularities. Many models was trained on this approach. The best model predicts a user's location, using as a training set only the places of users' tweets. The location of a person's can also be predicted using the activity of

their friends' network, in Facebook. This method first analyze friendship as a function of distance and rank and generate several observations regarding the interplay of geography and friendship. They find that at medium to long-range distances, the probability of friendship is roughly proportional to the inverse of distance. Then they present an algorithm to predict the physical location of a user, given the known location of her friends. The most interesting part is the comparison of the state-of-the-art techniques in the geolocation prediction. The authors considered that the results was under controversial conditions in order to make right conclusions, so they tried a comparison among them. To conclude since the data mining regarding the social media is at a primary stage, further work must be done in order to provide more accurate methods and techniques in the future.

6. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[2] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 11–18, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

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

[4] N. N. Bora. Summarizing public opinions in tweets. *International Journal of Computational Linguistics and Applications*, 3(1):41–55, 2012.

[5] R. O. Bueno, A. F. Bruzón, Y. Gutiérrez-Vázquez, and A. Montoyo. Ssa-uo: Unsupervised sentiment analysis in twitter. In *SemEval@NAACL-HLT*, 2013.

[6] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emerging topic detection based on a term aging model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):7, 2013.

[7] D. Chatzakou, V. Koutsonikola, A. Vakali, and K. Kafetsios. Micro-blogging content analysis via emotionally-driven clustering. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 375–380. IEEE, 2013.

[8] D. Chatzakou, N. Passalis, and A. Vakali. Multispot: Spotting sentiments with semantic aware multilevel cascaded analysis. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 337–350. Springer, 2015.

[9] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

[10] A. Cui, M. Zhang, Y. Liu, and S. Ma. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, pages 238–249, Berlin, Heidelberg, 2011. Springer-Verlag.

[11] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53, 2010.

[12] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[13] D. Gao, W. Li, and R. Zhang. Sequential summarization: A new application for timely updated twitter trending topics. In *ACL (2)*, pages 567–571. Citeseer, 2013.

[14] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Syst. Appl.*, 40(16):6266–6282, Nov. 2013.

[15] A. Giachanou and F. Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

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

[17] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[18] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. 2008.

[19] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM*, 13:273–282, 2013.

[20] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*, pages 188–197, 2015.

[21] V. Kharde, P. Sonawane, et al. Sentiment analysis of twitter data: A survey of techniques. *arXiv preprint arXiv:1601.06971*, 2016.

[22] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 459–464, New York, NY, USA, 2012. ACM.

[23] L. Kong, Z. Liu, and Y. Huang. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13):1681–1684, 2014.

[24] R. Li, S. Wang, and K. C.-C. Chang. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, 5(11):1603–1614, 2012.

[25] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.

[26] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users.

ICWSM, 12:511–514, 2012.

[27] J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 459–468. ACM, 2013.

[28] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[29] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[30] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[31] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.

[32] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, 2009.

[33] A. Severyn and A. Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 464–469, 2015.

[34] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. 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, pages 53–63, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[35] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565, 2014.

[36] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, New York, NY, USA, 2011. ACM.

[37] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. 2011.