Aristotle University of Thessaloniki
Department of Informatics

Twitter Stream Mining: An analysis of basic techniques

Pardalidou Christina UID: 690
Tsiliggiris Alexandros UID: 703
Karapostolakis Sotiris UID: 666

Contents

- Introduction
- Event Detection
- Sentiment Analysis
- Geo-Location
- Conclusions
- References

Introduction

- World Wide Web → Dynamic Information Resource
- Social Networks
  - Opinion Expression
  - Broadcasting News
  - Communication
- Twitter is the most popular micro-blogging service
- Monitoring and analysing → Significant Information
  - Users can be instantly informed about event/news
  - Companies use Twitter for advertising and promotion reasons or build reputation; analysing users' sentiment.

Introduction

- Twitter is not only endless database of short messages →

Event Detection

  - Quick Information Delivery
  - Optimal query retrieval in search engines

- User-Generated-Content(UGC) has led many companies to Sentiment Analysis.
- Reasons for Sentiment Analysis
  - Better proposals from recommendation systems
  - Sites for creation of summaries from people's reviews
  - Companies

Introduction

- Strong connection between **Event Detection** and **Location**
- Users are not only individuals, but part of user groups
- Geo-Location Prediction
  - User-located news
  - Regional Advertisement
- **Main problem**: Only 1% of tweets are geo-tagged.

Event Detection

- Goal: Extract an event by analysing tweets
- Why it isn’t simple
  - Large data stream
  - Must be processed online
  - Noise
    - Tweets about music, movies, etc.

A detection approach

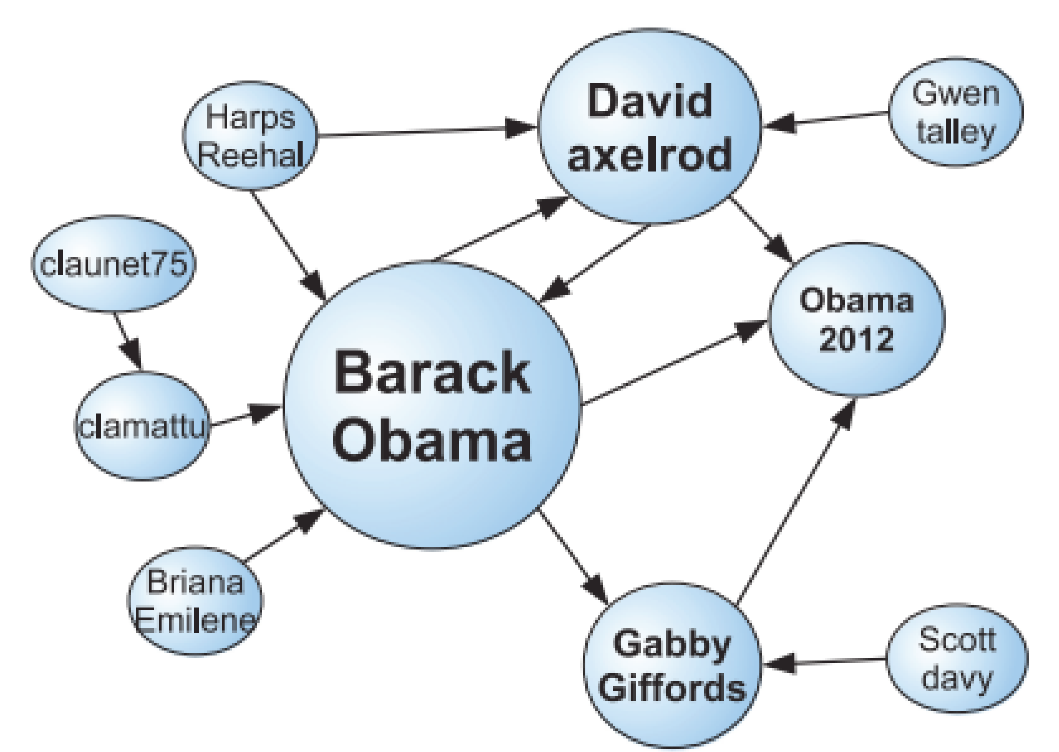
- Based on the idea that:
  - Users tweet about a particular event
  - Events are reported in real time
- Online algorithms must be used
  - To capture the “pulse of the moment”

Difference from newswire

- News reporters are not known
  - They appear randomly (random users)
- News reporters must be identified

Identifying users

- Users with a high amount of followers
  - Usually post news
  - Are influential in the network



Approaches: Identifying news tweets

- By training a Naive Bayes classifier
    - Tweets from influential users are usually news
- Term frequency and sorting
- Online clustering
- Living organism

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). Sentiment analysis involves classifying opinions in text into categories like "*positive*" or "*negative*" or "*neutral*".

- **Twitter Sentiment Analysis** (*TSA*) tackles the problem of analyzing the messages posted on Twitter in terms of the sentiments they express.

Sentiment Analysis

- 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.
  - SA at the **sentence** level aims to classify each sentence as positive or negative.
  - SA at the **entity** level detects the sentiment polarity of a specific entity/target of a particular object.

- 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*.

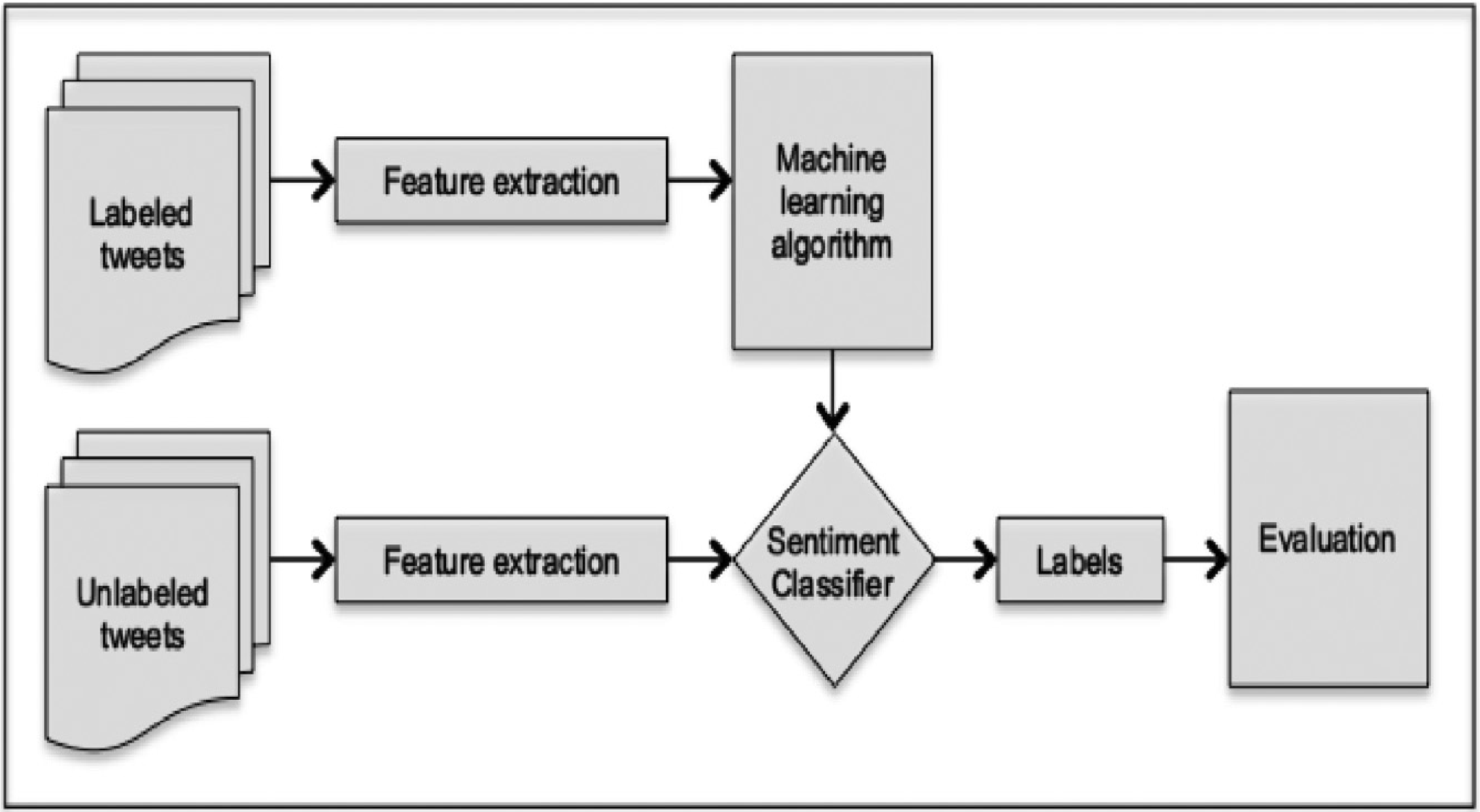
Sentiment Analysis

- Four different classes can be identified in the literature of TSA:
  - **Machine Learning**
  - **Lexicon-Based**
  - **Hybrid** (*Machine Learning & Lexicon-Based*)
  - **Graph-Based**

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:
  - **Naive Bayes** (*NB*)
  - **Maximum Entropy** (*MaxEnt*)
  - **Support Vector Machines** (*SVM*)
  - **Logistic Regression** (*LR*)
  - **Random Forest** (*RF*)
  - **Conditional Random Field** (*CRF*)

Machine Learning



Typical process for sentiment classification.

Machine Learning

- **Bakliwal et al. 2012** 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. The best results were achieved by the combination of NLP and Twitter-specific features.

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 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 (e.g *yolo, gr8*) but also has a dynamic nature with new expressions and hashtags emerging from time to time.

Lexicon-Based

- **Ortega et al. 2013** 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. However, the authors did not compare their method with existing techniques to prove its effectiveness.

Hybrid

- An interesting hybrid method was presented by **Ghiassi et al. 2013**, 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.

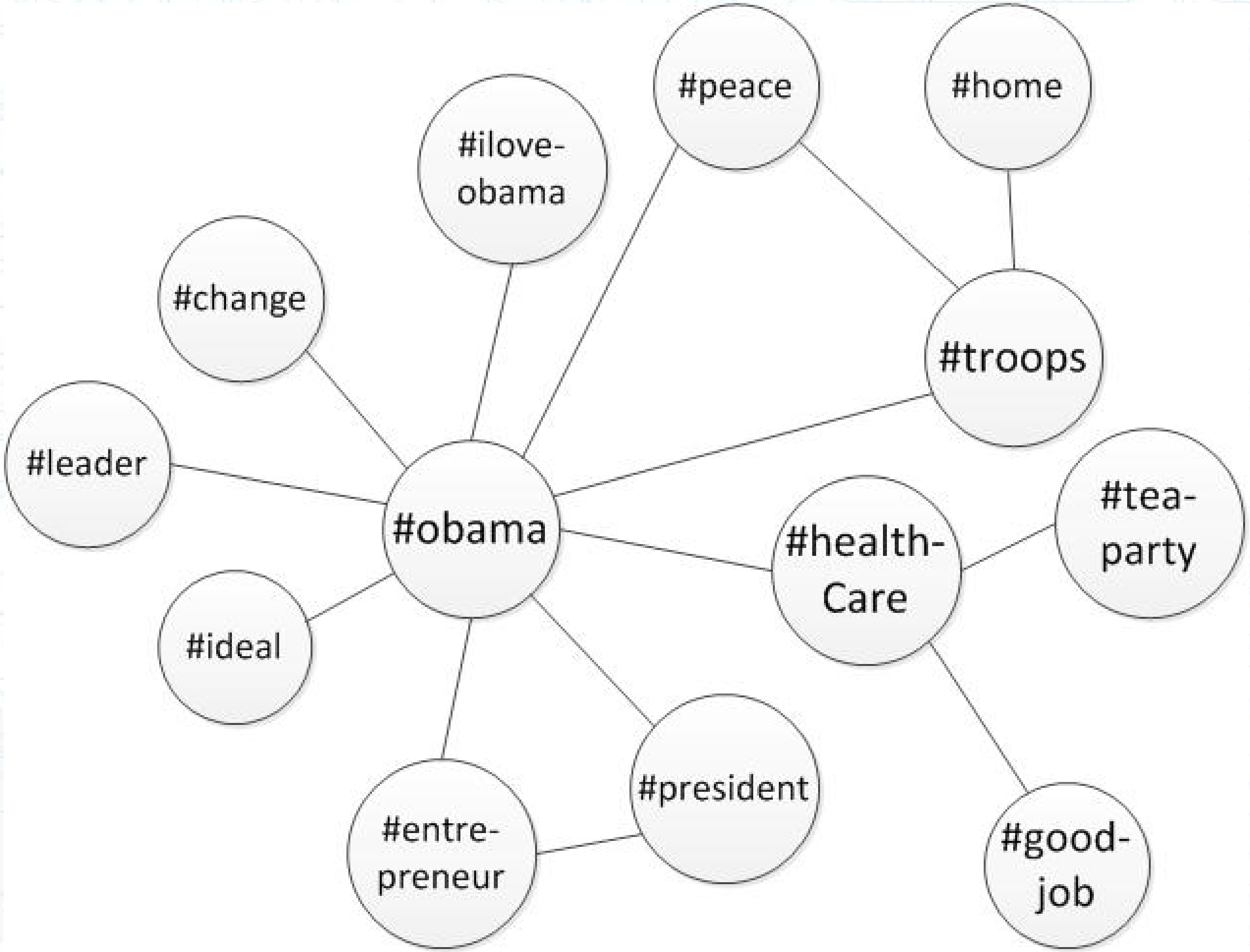
Graph-Based

- 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.

Graph-Based

- Some of the first to apply a label propagation method for TSA were **Speriosu et al. 2011**. 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.

Graph-Based



An example of a Hashtag Graph Model

Geolocation

Geolocation at Different Granularities

→ Prediction at city-level

→ 1,524,522 tweets from 9551 users

→ Google's geocoding API, Twitter's Rest API and Twitter's Streaming API

- Content-based statistical classifiers
  - Word Classifier
  - Hashtag Classifier
  - Place Classifier

Weka's Naive Bayes Multinomial as learning algorithm.

Geolocation

- Content-based heuristics classifiers
  - Local Place Heuristic Classifier
  - Visit Place Heuristic Classifier
- Ensemble of Location Classifiers

→ Classification Strength for a user is the inverse of the number of matching locations in the matching location distribution.

→ 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.

Geolocation

Comparison of the state-of-the-art techniques

→ Evaluation criteria

- Accuracy of a post's location.
- Accuracy of a user's all posts location.
- Maximizing the posts whose location is able to be inferred.
- Metrics: AUC, Post Coverage and Median-Max.

→ Experiments

- Cross Validation, Self-Reported locations and Temporal Decay

Geolocation

- Using self-reported locations as ground truth yields far fewer ground truth locations than reported in earlier studies and even in the best configuration for matching these locations to coordinates, systems trained on self-reported locations suffer a large drop in accuracy, with the Median-Max error doubling for all but the lowest- performing systems.
- According to, Geoinference methods that predict locations of future posts, it is concluded that while accuracy rates only decrease slightly, Post Coverage was estimated to drop by half after a four month period.

Conclusions

- There is a need for designing of better feature extraction.
- More accurate filtering and detection algorithms.
- Improved techniques to combine and analyze information from multiple sources and multiple languages.
- Methods from the machine-learning field are applied for TSA on a more frequent rate compared to the rest of the approaches.
- SVM and NB classifiers are the most prevalent.

Conclusions

- One of the most reliable and straightforward approaches to geolocation prediction is IP-based methods, until now.

- Some of state of the art approaches′ results regarding to geolocation are controversial and need to be examined.

- Some approaches show that the location of a person can also be predicted using the activity of their friends′ network.

References

S. Petrovi'c, 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.

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

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.

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.

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.

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.

R. O. Bueno, A. F. Bruzon, Y. Gutierrez-Vazquez, and A. Montoyo. Ssa-uo: Unsupervised sentiment analysis in twitter. In SemEval@NAACL-HLT, 2013

References

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

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.

J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users.ICWSM, 12:511–514, 2012.