Twitter Sentiment Analysis

Instructor: Ekpe Okorafor

1. Big Data Academy - Accenture
2. Computer Science - African University of Science & Technology

Ekpe Okorafor PhD

Affiliations:

- **Accenture – Big Data Academy**
  - ❑ Senior Principal & Faculty, Applied Intelligence
- **African University of Science & Technology**
  - ❑ Visiting Professor, Computer Science / Data Science
  - ❑ Research Professor - High Performance Computing Center of Excellence

Research Interests:

- Big Data, Predictive & Adaptive Analytics
- Artificial Intelligence, Machine Learning
- Performance Modelling and Analysis
- Information Assurance and Cybersecurity.
- High Performance Computing & Network Architectures
- Distributed Storage & Processing
- Massively Parallel Processing & Programming
- Fault-tolerant Systems

Email: ekpe.okorafor@gmail.com; eokorafo@ictp.it; eokorafor@aust.edu.ng
Twitter: @EkpeOkorafor; @Radicube

Agenda

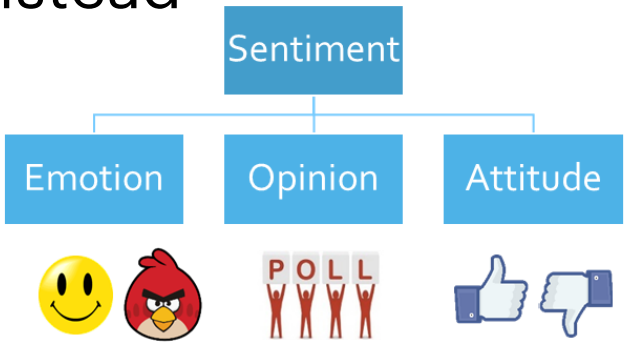
- Introduction
- Twitter Sentiment Analysis
- Use Cases

Agenda

- Introduction
- Twitter Sentiment Analysis
- Use Cases

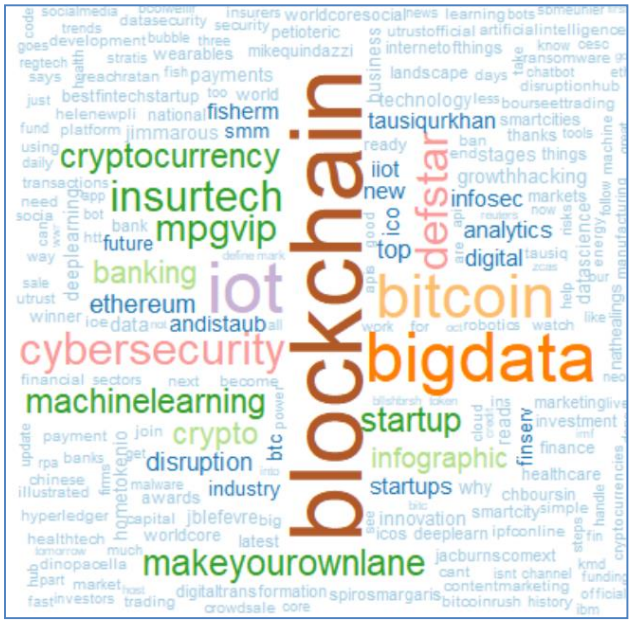
Terms

- Sentiment
  - A thought, view, or attitude, especially one based mainly on emotion instead of reason



- Sentiment Analysis
  - aka opinion mining
  - use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text

Motivation



This is by no means exhaustive!

- ❑ Consumer information
  - Product reviews
- ❑ Marketing
  - Consumer attitudes
  - Trends
- ❑ Politics
  - Politicians want to know voters' views
  - Voters want to know politicians' stances and who else supports them
- ❑ Social
  - Find like-minded individuals or communities

Problem

- Which features to use?
  - Words (unigrams)
  - Phrases/n-grams
  - Sentences
- How to interpret features for sentiment detection?
  - Bag of words (IR)
  - Annotated lexicons (WordNet, SentiWordNet)
  - Syntactic patterns
  - Paragraph structure

Challenges

- Harder than topical classification, with which bag of words features perform well
- Must consider other features due to…
  - Subtlety of sentiment expression
    - irony
    - expression of sentiment using neutral words
  - Domain/context dependence
    - words/phrases can mean different things in different contexts and domains
  - Effect of syntax on semantics

Approaches

- Machine learning
  - Naïve Bayes
  - Maximum Entropy Classifier
  - SVM
  - Markov Blanket Classifier
    - Accounts for conditional feature dependencies
    - Allowed reduction of discriminating features from thousands of words to about 20 (movie review domain)

Assume pairwise independent features

- Lexicon-based
  - Dictionary
  - Corpus

- Hybrid

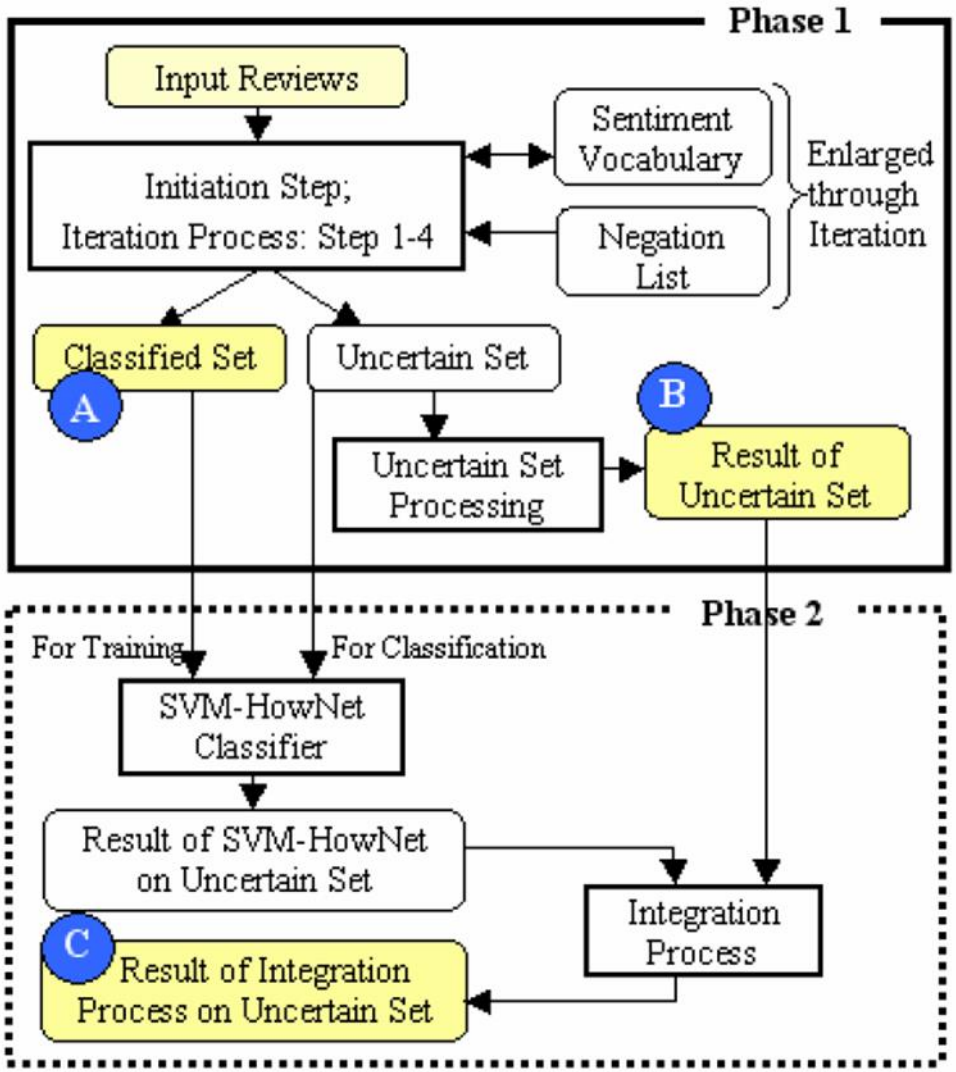
Machine Learning Approach

- Advantages:
  - Tend to attain good predictive accuracy
    - Assuming you avoid the typical ML mishaps (e.g., over/under-fitting)
- Disadvantages:
  - Need for training corpus
    - Solution: automated extraction (e.g., Amazon reviews, Rotten Tomatoes) or crowdsourcing the annotation process (e.g., Mechanical Turk)
  - Domain sensitivity
    - Trained models are well-fitted to particular product category (e.g., electronics) but underperform if applied to other categories (e.g., movies)
    - Solution: train a lot of domain-specific models or apply *domain-adaptation* techniques
    - Particularly for Opinion Retrieval, you'll also need to identify the domain of the query!
  - Often difficult/impossible to rationalize prediction output

Lexicon Based Approach

- Advantages:
  - Can be fairly accurate independent of environment
  - No need for training corpus
  - Can be easily extended to new domains with additional affective words
    - e.g., “amazeballs”
  - Can be easy to rationalise prediction output
  - More often used in Opinion Retrieval (in TREC, at least!)
- Disadvantages:
  - Compared to a well-trained, in-domain ML model they typically underperform
  - Sensitive to affective dictionary coverage

Hybrid Approach



Agenda

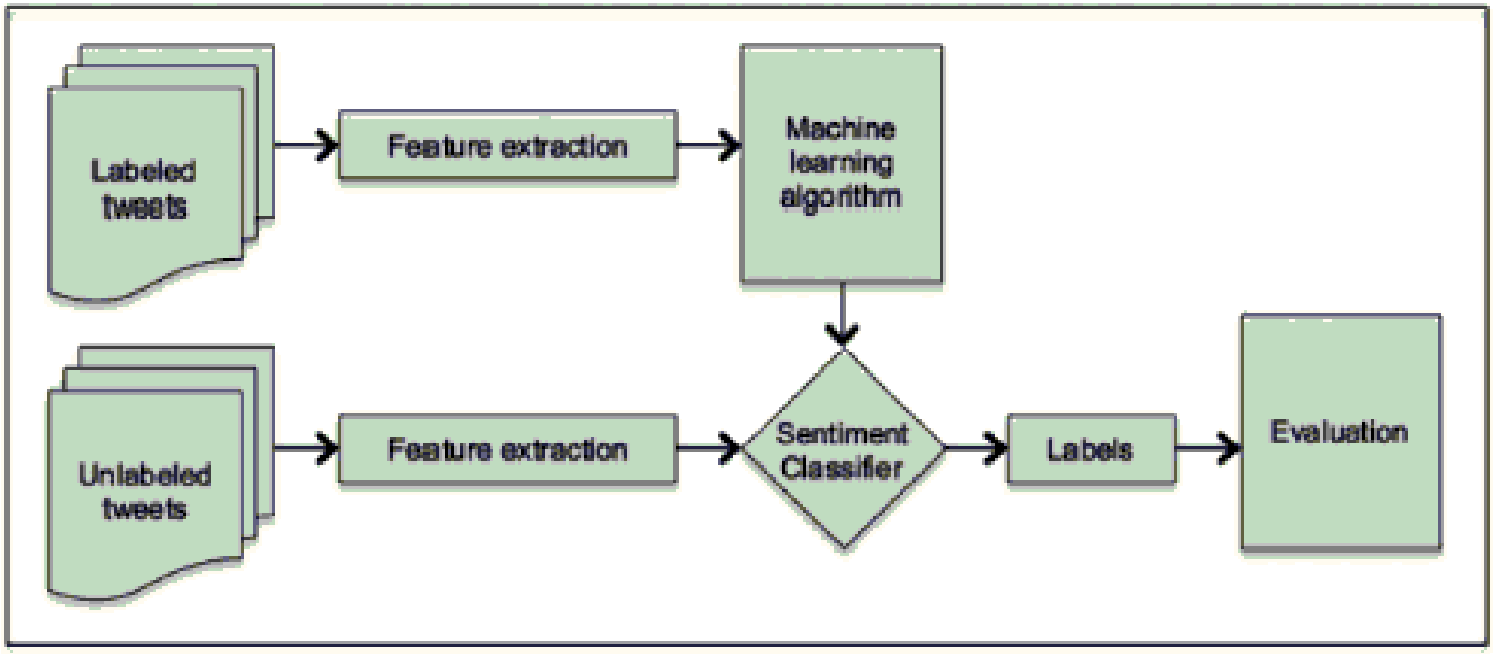
- Introduction
- Twitter Sentiment Analysis
- Use Cases

Introduction

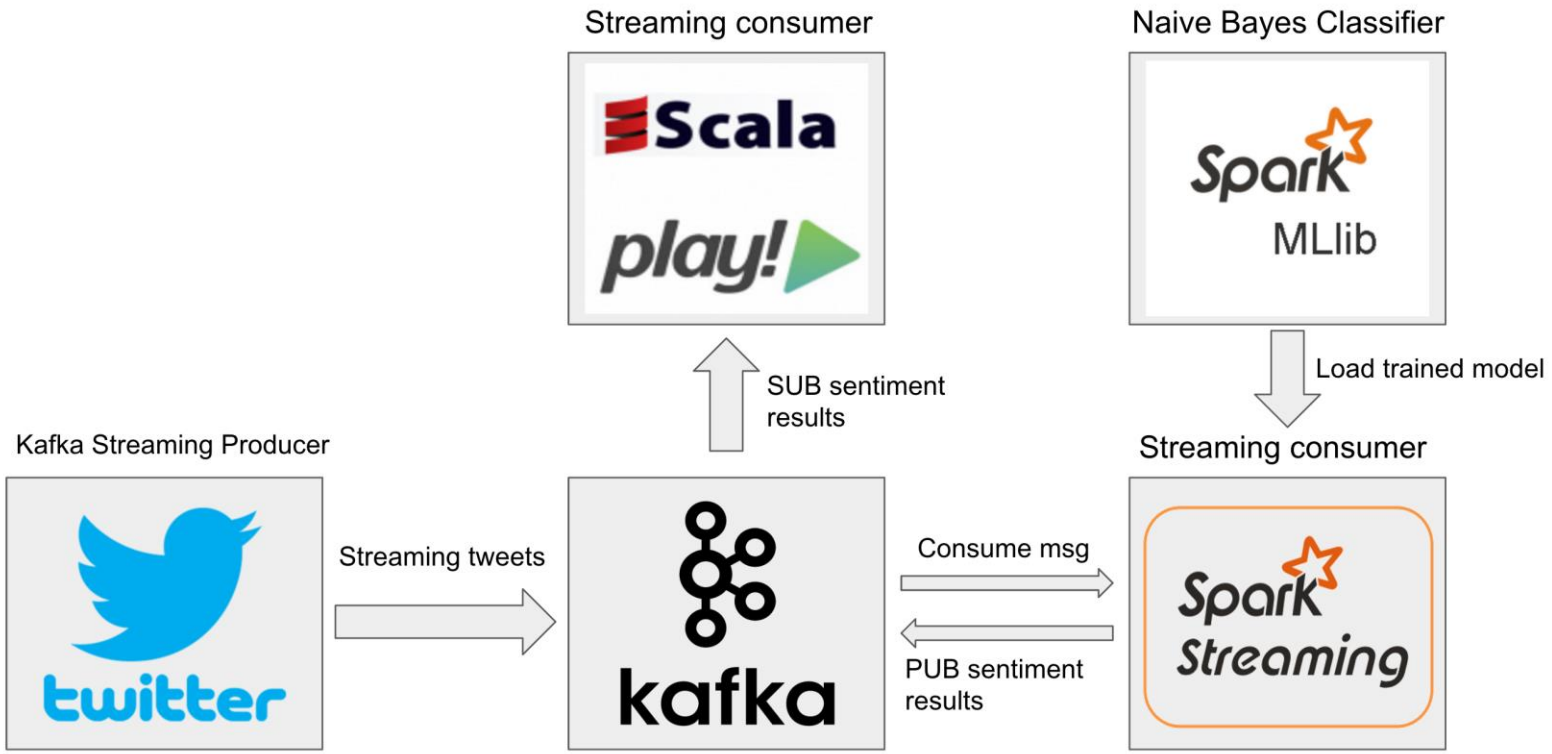
- Social Media
  - User-generated content
  - Research Areas
    - Opinion Mining (OM) – subjectivity analysis
    - Sentiment Analysis (SA) – sentiment polarity detection
- Twitter
  - Popular microblog
  - Opinions on various topics
- Twitter Sentiment Analysis (TSA)
  - Analyze messages posted on Twitter
  - Short length
  - Informal type

Introduction

- The majority of TSA methods use a method from the field of machine learning, known as classifier.



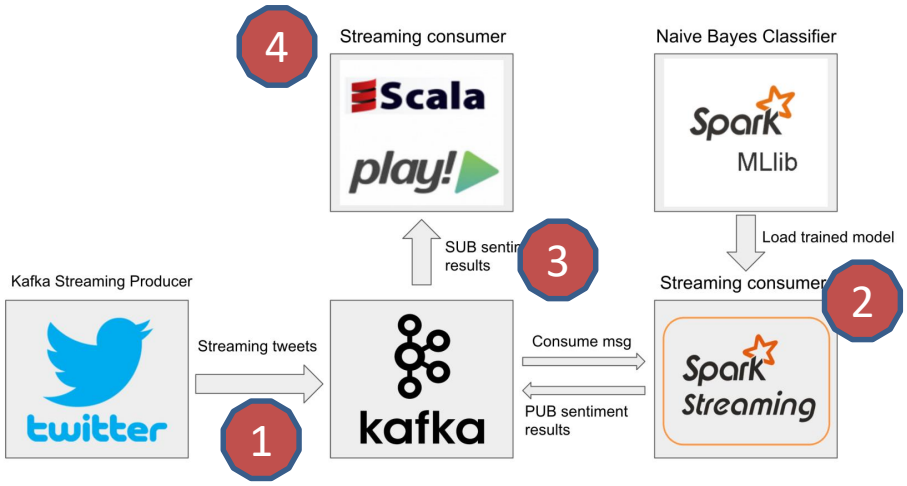
Implementation - Architecture



Modules

- ❑ Kafka twitter streaming producer
- ❑ Sentiment analysis consumer
- ❑ Scala play server consumer

Data Flow



1. Kafka twitter streaming producer publishes streaming tweets on the 'tweets' topic to the central **Apache Kafka**, and sentiment analysis consumer has subscribed that 'tweets' topic.
2. The sentiment analysis consumer leverage Apache **Spark Streaming** to perform batch processing on incoming tweets and load trained **Naive Bayes model** to perform sentiment analysis.
3. And then accumulated count of each **positive sentiment and negative** sentiment reduced by each location are published on topic 'sentiment' to central Kafka, and this 'sentiment' topic subscribed by Scala Play Server.
4. The sentiment analysis results will be send to web clients through webSocket connections.

Machine Learning - Classifier

Bayes' theorem describes the probability of an event, based on conditions that might be related to the event:

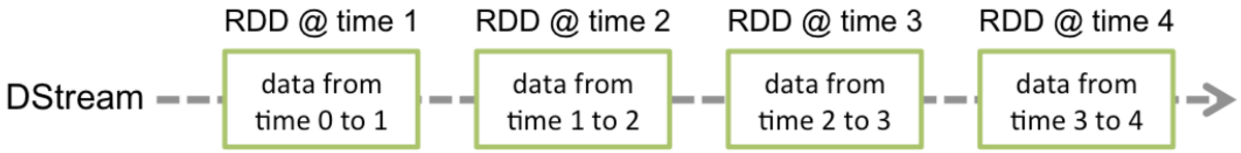
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Naive Bayes - family of probabilistic **classifiers** of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features solving classification problem.
- **Apache Spark MLlib** supports Multinomial Naive Bayes and Bernoulli Naive Bayes.

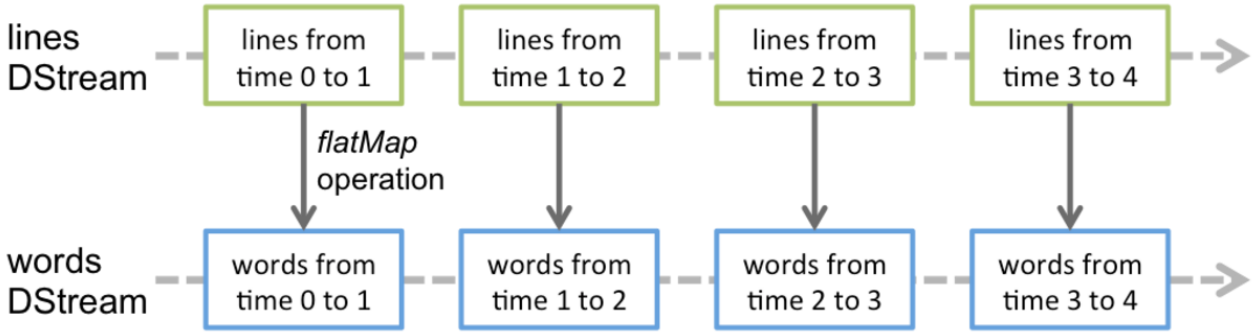
Real Time Streaming – Spark Streaming

Spark Streaming

- Spark streaming leverages spark core to perform streaming analysis. Discretized Stream or DStream is the basic abstraction provided by Spark Streaming.



- Each RDD in a DStream contains data from a certain interval
- Any operation applied on a DStream translates to operations on the underlying RDDs.



Agenda

- Introduction
- Twitter Sentiment Analysis
- Use Cases

Use Cases – Public Health

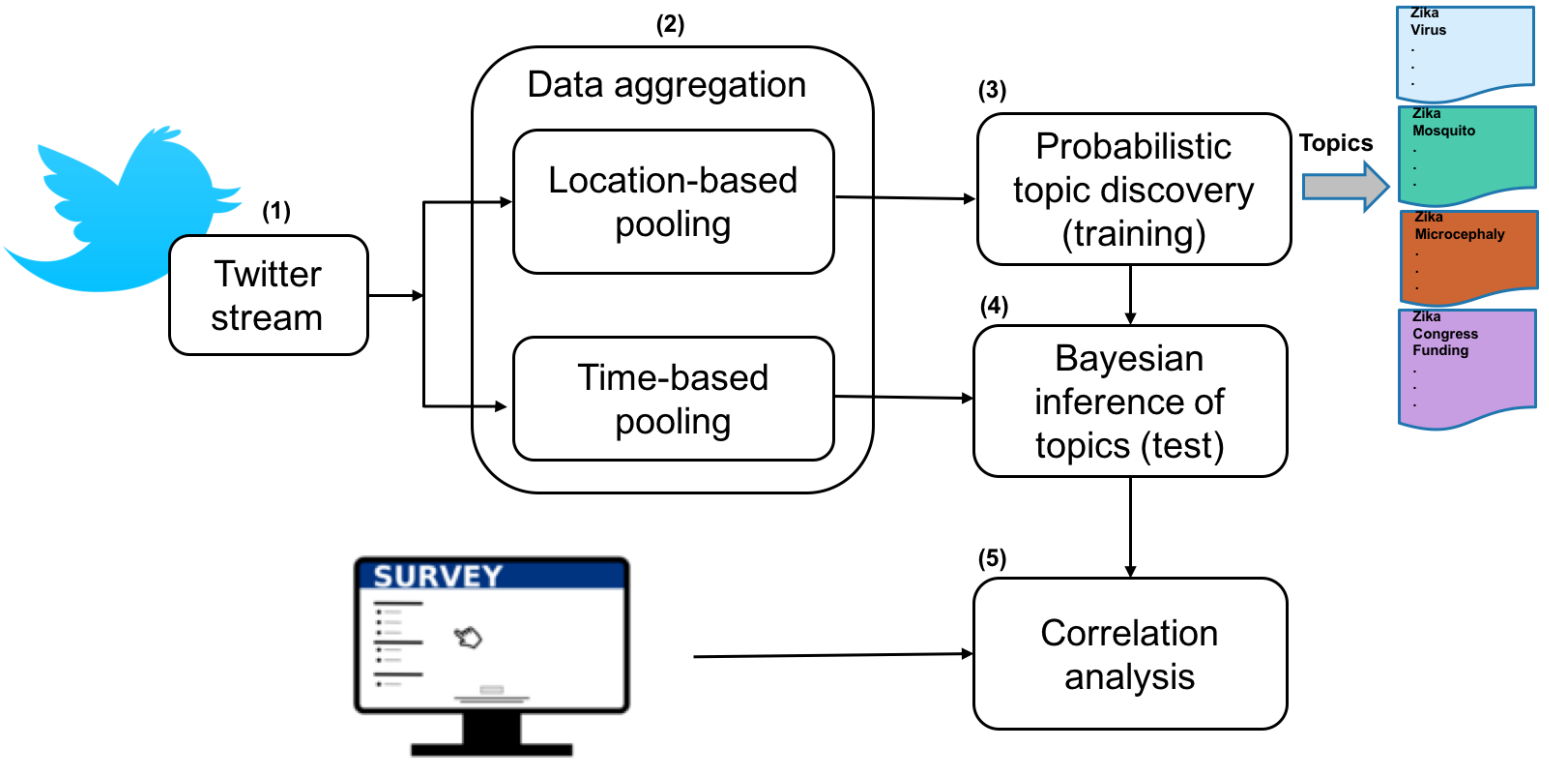
Associations of Topics of Discussion on Twitter With Survey Measures of Attitudes, Knowledge, and Behaviors Related to Zika: Probabilistic Study in the United States

Mohsen Farhadloo[1,2], PhD ; Kenneth Winneg[2], PhD ; Man-Pui Sally Chan[1], PhD ;

Kathleen Hall Jamieson[2], PhD ; Dolores Albarracin[1], PhD

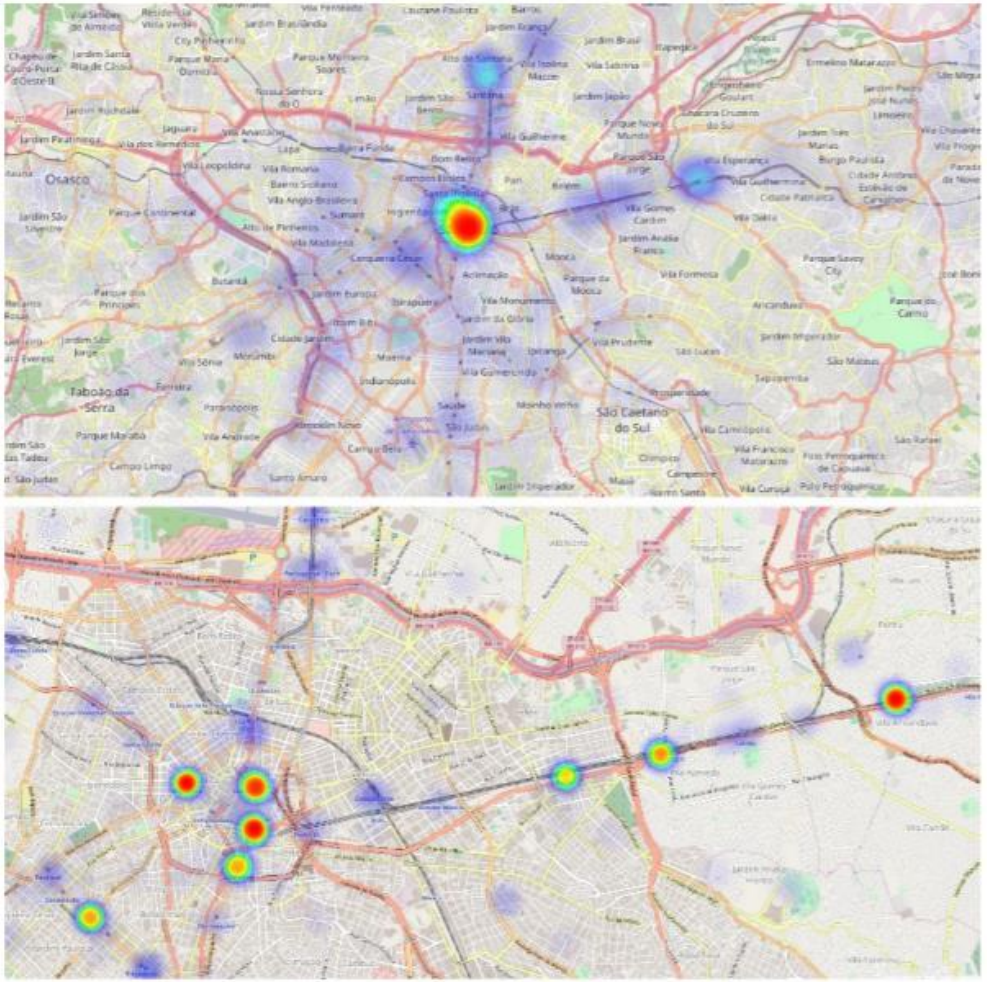
[1]University of Illinois at Urbana-Champaign, Champaign, IL, United States

[2]Annenberg Public Policy Center, University of Pennsylvania, Philadelphia, PA, United States



Use Cases – Smart Cities

- Governments across the world are trying to move closer to their citizens for better smart city monitoring and governance.
- Twitter Sentiment Analysis is opening new opportunities to achieve it.

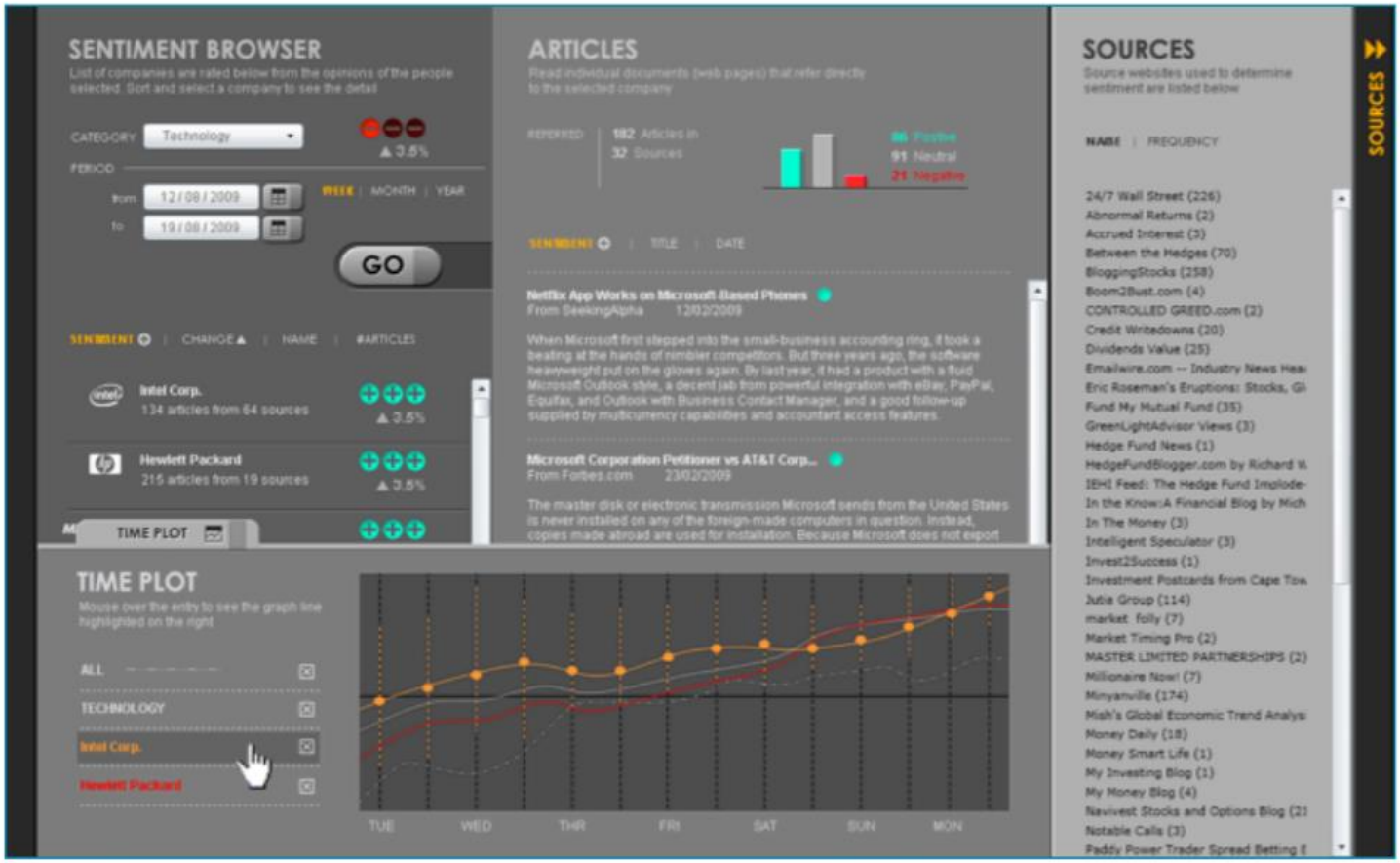


Heat map of city to positive tweets

Use Cases – Real Time Political Analysis

- Data-driven media and journalism
- PR management for political figures and parties

Use Cases – Financial Analysis



Intelligent tools for aiding decision-making for financial traders and analysts

Use Cases – Radicalization Detection

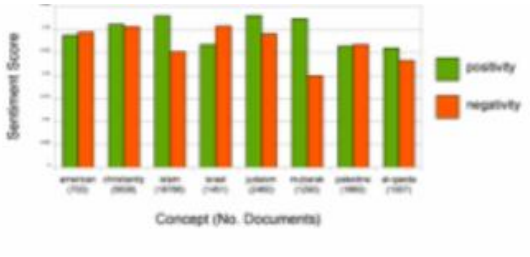


Figure 1. Sentiment Analysis Results: Males

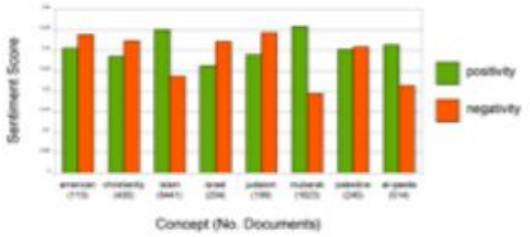


Figure 2. Sentiment Analysis Results: Females

Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation

Adam Bermingham[1], Maura Conway[2], Lisa McInerney[2],
Neil O'Hare[1], Alan F. Smeaton[1]
[1]CLARITY: Centre for Sensor Web Technologies
and [2]School of Law and Government,
Dublin City University, Glasnevin, Dublin 9, Ireland.

Abstract

The increased online presence of jihadists has raised the possibility of individuals being radicalised via the Internet. To date, the study of violent radicalisation has focused on dedicated jihadist websites and forums. This may not be the ideal starting point for such research, as participants in these venues may be described as "already made-up minds". Crawling a global social networking platform, such as YouTube, on the other hand, has the potential to unearth content and interaction aimed at radicalisation of those with little or no apparent prior interest in violent jihadism. This research explores whether such an approach is indeed fruitful. We collected a large dataset from a group within YouTube that we identified as potentially having a radicalising agenda. We analysed this data using social network analysis and sentiment analysis tools, examining the be described as "already made-up minds". In crawling YouTube[1] we can look at interaction aimed at those with little or no prior interest in violent jihadism. The present paper builds on previous research on the links between jihadi video and online radicalisation, and the contribution is a detailed analysis of a real YouTube dataset. This analysis uses an application of sentiment-, lexical- and social network analysis, which allows us to examine and characterise the users of radicalised forums, with particular emphasis on gender-based differences between users.

2. Related Work

In the sub-sections below we introduce related research into online radicalisation, followed by a summary of a previous case study, in which we identified the YouTube group

Sentiment analysis with social network analysis and automatic demographic profiling


ANY
QUESTIONS
?