Monitoring and Detecting Online Hate Speech from Twitter

MANDOLA

George Pallis
Assistant Professor, University of Cyprus
gpallis@cs.ucy.ac.cy
https://www.cs.ucy.ac.cy/~gpallis/
Barack Obama: Now Is the Greatest Time to Be Alive

That’s how we will overcome the challenges we face: by unleashing the power of all of us... Not just making our social networks more fun for sharing memes but also harnessing their power to counter terrorist ideologies and online hate speech.

Wired, November issue, 2016

https://www.wired.com/2016/10/president-obama-guest-edits-wired-essay/
Vint Cerf: This is a sociological problem in large measure

Q: Hacking, Hate Speech, so-called fake news - what do you suggest, so that these phenomena do not overwhelm?

“If there are technologies with which we can prevent abuse, we must apply them.... There are no simple algorithms with which to identify content, which are falsified or illegal. To recognize this is a particularly tricky problem....

However, everyone must also take care of their own security. This also applies to me as someone who works at Google. My job is to provide you with tools to help you ensure your safety. But if you do not use it and your account is cracked, you can not blame Google for it.” (Wired, March 2017)
Fights against Hate Speech

• Germany’s justice minister Heiko Maas proposed fining social media up to €50m for not responding quickly enough to reports of illegal content or hate speech. (March 2017)

• The law would require social media platforms to come up with ways to make it easy for users to report hateful content. Companies would have 24 hours to respond to “obviously criminal content” or a week for more ambiguous cases.
What is hate speech?

- In the Oxford dictionary hate is defined as an emotion of "intense dislike", for someone or something, of an "aversion to" something.
- This feeling of "hatred or intolerance" can be verbalised in a speech (becoming this way a "hate speech") and used to express "hatred or intolerance of other social groups, especially on the basis of race or sexuality".
- Since there is no universal definition, some web pages and social platforms have adopted their own definitions of hate speech.
Hate speech and Internet Giants

- **Facebook** defines the term "hate speech" as "direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease".
- **Twitter** does not provide its own definition, but simply forbids to “publish or post direct, specific threats of violence against others.”
- **YouTube** website clearly says it does not permit hate speech, which it defines as “speech which attacks or demeans a group based on race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity.”
- **Google** makes a special mention on hate speech in its User Content and Conduct Policy: “Do not distribute content that promotes hatred or violence towards groups of people based on their race or ethnic origin, religion, disability, gender, age, veteran status, or sexual orientation/gender identity.”
Hate speech and Internet Giants

• Monika Bickert, head of global policy management at Facebook said: “There’s no place for hate speech on Facebook. We urge people to use our reporting tools if they find content that they believe violates our standards so we can investigate. Our teams around the world review these reports around the clock and take swift action.”

• Twitter’s head of public policy for Europe, Karen White, said: “Hateful conduct has no place on Twitter and we will continue to tackle this issue head on alongside our partners in industry and civil society.”
Hate speech and Internet Giants

On May 2016, Facebook, Google and Twitter signed a code of conduct, announcing a set of standards for dealing with hate speech, including:

- A promise to review the majority of reports of illegal hate speech and remove the offending content within 24 hours
- Making users aware about what is banned by each company
- Training staff to let them better spot and respond to online hate speech.

In practice the companies only reviewed 40 per cent of the recorded cases in less than 24 hours.
Hate Speech and Twitter

- Nick Pickles, head of public policy and government for Twitter UK, admitted that Twitter was doing a bad job at keeping abusive content off of its platform. "We don't communicate with the users enough when they report something, we don't keep people updated enough and we don't communicate back enough when we do take action”.

- Twitter Takes More Steps To Block Internet Trolls And Hate Speech (March 2017)
  - Twitter says that cannot rely only on user and government reports
  - A new "safe search" feature filters tweets with sensitive content, and those from blocked or muted accounts.
  - Twitter says it will ban Trump if he breaks hate-speech rules
@USERNAME absolute bastard!

No hate speech
These faggot ass niggas really think they a female

Hate speech
Mandola Project Objectives

- **Monitor** the spread of online hate-related speech.
- **Analyse** its content and the **categories** to which it might belong (Ethnicity, Nationality, Religion, Gender, Sexual, Class, Disability)
- **Store** and **visualize** actionable information for **policy makers**, to promote policies against online hate speech, and **citizens**, to raise their awareness
- Do that **without** holding any user’s sensitive data, by processing data on the fly, following a procedure approved by the Cyprus Data Protection Commissioner

Article 7(1)(2) of the Personal Data Protection, N. 138(I)/2001

http://mandola-project.eu/
MANDOLA Tools

• A bundle of tools to tackle online hate-speech and clear our objectives

Monitoring Dashboard

Mobile Application
Monitoring Dashboard

- **Consists of the following modules:**
  - Data Streams Collection (Twitter and Web pages)
  - Hate-speech Data Analysis (Multi-lingual Classifier)
  - Data Storage (Hate-speech statistical database)
  - MANDOLA Application Interface – API
  - Monitoring Dashboard Web Application
Data Stream Collections

- Consist of two sub-modules:
  - **Twitter data stream** is collected via Twitter Stream API using a framework to efficiently retrieve large collection of tweets per day (developed by University of Cyprus)
  - **Web data stream** is collected via meta-search engine that crawls possible hate-related web pages (developed by University Autonomous of Madrid, Spain)

- Each collected stream is fed to Apache Kafka, a queuing system for supporting streaming data processing

“Distributed Large-Scale Data Collection in Online Social Networks” H. Efstathiades et al. IEEE CIC 2016
Twitter Data Stream

“Distributed Large-Scale Data Collection in Online Social Networks” H. Efstathiades et al. IEEE CIC 2016
Web Data Stream

URL 1
- Extract Links
- Save Links

URL N
- Extract Links
- Save Links

Crawling Module
- Get Links
- Scrap Web
- Analyze & Extract Data
- Save New Links
Hate-speech Data Analysis

- Input data - comes from Apache Kafka (distributed streaming platform) - are preprocessed to enter the Classifier
- Multi-lingual Hate-speech Corpus, continuously enriched, supports and retrains the Classifier
- The Classifier receives the input and outputs the Hate-speech analysis to be stored

https://kafka.apache.org/
Data Preprocessing

• Remove non-linguistic features as:
  • URLs e.g. http://www.google.com
  • Hashtags e.g. #MANDOLA
  • Emails e.g. gpallis@cs.ucy.ac.cy
  • User Mentions e.g. @mandola_project

• Expand any abbreviation found
  • e.g. “thanx” = “thank you”

• Apply more linguistic processing

“Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus” Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong, Carolyn P. Rose, 2016
Hate-speech Classifier

- Multi-lingual classifier trained with machine learning techniques 🇬🇧 ‏🇫🇷 ‏🇮🇹 ⚽️ 🇪🇸 🇬🇷
- Trained with **MLlib**, Apache Spark's scalable machine learning library
- Training data and testing data comes from the Multi-lingual corpus
- The output consists of statistical data **only**, and does not hold vulnerable information

[MLlib](http://spark.apache.org/mllib/)
Spark Components – MLlib

- MLlib (Machine Learning Library) is a distributed machine learning framework above Spark
- Spark MLlib is nine times as fast as the Hadoop disk-based version of Apache Mahout
- Spark MLlib provides a variety of machine learning classic algorithms

MLlib
machine learning

Spark Streaming

train models
use trained model

with live data
Classification Algorithm

- **Class**: Hate or No Hate

- **Target Labels**: 7 categories of hate speech (ethnicity, nationality, religion, gender, sexual, disability, class)

Model

- In multiclass-multilabel classification, the goal is to assign one or more labels to each instance in an instance space
- Each label associates an instance with one of 2 possible classes (hate / no hate)
  - Stochastic Gradient Descent (SGD) Classifier supports multi-class classification by combining multiple binary classifiers in a “one versus all” (OVA) scheme.

Ofer Dekel, Ohad Shamir (Microsoft Research): Multiclass-Multilabel Classification with More Classes than Examples. AISTATS 2010: 137-144
Why Stochastic Gradient Descent (SGD)?

- SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in **text classification** and **natural language processing**
- Given that the data is sparse, the classifiers in this module easily scale to problems with more than $10^5$ training examples and more than $10^5$ features

The advantages of Stochastic Gradient Descent are:
- Efficiency
- Ease of implementation (lots of opportunities for code tuning)
Multi-lingual Corpus

- **Multi-lingual sets** of hate-related speech
  - One classifier per language
  - The Corpus is produced from the **Filtering Mechanism** and the **Social Scientists** who manually annotate text
  - The system continuously enriches the corpus

![Diagram](Diagram.png)
Hate Filtering Mechanism

- **Subjectivity Classifier:** Hate-speech tends to be opinionated, thus subjective sentences are more likely to contain hate-speech than objective. *(NLTK sentiment analysis - http://www.nltk.org/)*

- **Polarity Classifier:** Hate-speech tends to have negative meaning, thus sentences with negative polarity are more likely to contain hate-speech *(Hatebase and AFINN)*

References:

Social Scientists

• Their task is to annotate if provided data contain hate-speech and the category to which they belong.
• To do so the MANDOLA annotation system was developed.
Crowd-sourced analysis: issues of reliability and validity

- Use the Cohen's kappa statistic test which measures inter-rater agreement for qualitative (categorical) items
- Takes into account the possibility of the agreement occurring by chance

Further research

- Evaluate a quantitative approach to investigate rater effects for the classification of data
Monitoring Dashboard
Web application

- Built with **HTML5, CSS3** and **JavaScript**.
- Compatible with any web browser
- Responsive (**Twitter’s Bootstrap**) and compatible with any mobile device
  - Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive projects on the Web. Download Bootstrap at http://getbootstrap.com
- Can be found at: [http://mandola.grid.ucy.ac.cy:3000/](http://mandola.grid.ucy.ac.cy:3000/)
- Password: m@nd0la_20!6
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.
Dashboard – Hotspot

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL:
http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_2016
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.
Dashboard – Hotspot

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL: http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_2016
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL: http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_2016
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL: http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_2016
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.
Dashboard – Statistics
Language Usage

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL:
http://mandola.grid.ucy.ac.cy:3000/

Password: m@nd0la_2016
Dashboard – Statistics
Hate Percentage per Category

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL:
http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_2016
Dashboard – Statistics
Hate Percentage per Country

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL:
http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_20!6
Dashboard – Statistics
Top Three Countries

These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.

URL: http://mandola.grid.ucy.ac.cy:3000/
Password: m@nd0la_20!6
These Visualizations are based on mock data and have nothing to do with real hate-speech analysis.
Ongoing Development

- Event labeling in hate-speech timeline
  - Users will be able to add events in the timeline in order to relate them with the chart
  - Identify correlations between the hate-speech analysis and certain events
  - Provide animation of hate-speech analysis during a period of time
  - Help the users to see the changes of hate-speech via animating the results for a certain duration
Mobile Application

• Its main objectives are to:
  • Support anonymous reporting
  • Raise online hate-speech awareness by viewing statistical analysis and FAQs
  • Compatible for Android and iOS devices

• Development technologies:
  • **Cordova** enables wrapping up of CSS, HTML, and JavaScript code depending upon the platform of the device

https://cordova.apache.org/
Mobile Application Interfaces

- **Awareness View**
  - HATE-SPEECH ENCOUNTERS
    - 4 out of 5 encounter online hate-speech.
    - 2 out of 5 have personally felt attacked.
  - GLOBAL HATE-SPEECH
    - 2014: 30%
    - 2015: 55%
    - 2016: 70%

- **FAQs View**
  - FAQs FREQUENTLY ASKED QUESTIONS
    - Search FAQs...
    - What is Hate Speech?
    - What is legally considered Hate Speech?
    - Who are the victims of hate speech?
    - I believe I have encountered hate speech. What can I do about it?
    - What does Freedom of Expression allow us to say?
    - Why words can be harmful and sometimes dangerous?

- **Report View**
  - No title
    - It's now easier than ever for orga...
    - 4 hours ago
  - No title
    - The biggest news agency from China...
    - 18 hours ago
  - No title
    - Now in its thirteenth season, Kimm...
    - 18 hours ago

- **Settings View**
  - Default OCR language
    - English
  - Manage languages for OCR
    - Install or uninstall languages for the OCR
  - Hatespeech analysis
    - Perform analysis on report
  - Keep image data
    - Keep cropped images from OCR
  - Enable MANDOLA bubble
    - MANDOLA background mode
Mobile Application

• Reporting:
  • The user is required to fill the source URL, the hate-speech text and the categories to which it belongs
    • Optionally can fill a title input for more details
  • The user must be able to report hate-speech encounters from both public and private sources.
  • Thus we provide two main methods for reporting:
    • Using MANDOLA Proxy server and custom browser to report public hate-speech encounters
    • Using Optical Character Recognition – OCR to report private hate-speech encounters
Mandola Reporting Application architectural diagram
Optical Character Recognition Module

• The Tesseract OCR system is used (Apache license, sponsored by Google)
• Recognizes a total of 100 languages
• Available for Linux, Windows and Mac OS X

https://github.com/tesseract-ocr
Mandola Proxy Module

- It utilizes the Wayback Machine
  - Enables users to see archived versions of Web pages across time.
  - The implementation of the MANDOLA Proxy was done by PyWb, a Python implementation of the Wayback Machine.
  - TextHighlighter, a JavaScript library that enables website text highlight, is used.

![Diagram of Mandola Proxy Module](https://archive.org/web/)
Mobile Application

• **MANDOLA Bubble:**
  • It is a background process for the automation of the form filling after the user copies the URL
  • Coming soon in Android and iOS Market!
Mandola Bubble

START SERVICE

MANDOLA BUBBLE SERVICE

LISTENS TO URL COPY EVENT

IF BUBBLE PRESSED GENERATE REPORT
Acknowledgments

This work has been produced with the financial support of the Rights, Equality and Citizenship (REC) Programme of the European Union under Grant Agreement JUST/2014/RRAC/AG/HATE/6652

Prof. Marios Dikaiakos
Assoc. Prof. Tatiana Synodinou

Software Engineers: Demetris Paschalides, Demos Stefanides

URL: http://mandola.grid.ucy.ac.cy:3000/Password: m@nd0la_2016