But I did not Mean It!- Intent Classification of Racist Posts on Tumblr

Swati Agarwal*, Ashish Sureka**
*Indraprastha Institute of Information Technology, Delhi, India
**ABB Corporate Research Centre, Bangalore, India

6th European Intelligence and Security Informatics Conference (EISIC)
Uppsala, Sweden
August 18, 2016
"Brothers and sisters: This religion is not religion of talk, it's not religion of showing off how much knowledge you have, this religion is about your willingness to sacrifice for Allah ﷺ.
Shaykh Anwar al-Awlaki

Stop Islam?! yeah, Stop Islam from being stereotyped and misunderstood by Islamophobic and ignorant racists.

♥ 5,851 notes {block:IfDisqusShortname}

Stop Islam stop islamophobia stop ignorance islamophobia islam brussels attack
“can we please wake up from this equality bullshit and just start getting rid of people like fags, blacks, jews, women, and anyone who cant provide for their country? we need to seriously do something before the world turns to shit”

If you can’t understand what it’s like to be a tiger in a cage, you’ll never understand what it’s like to be black in America.

#iloveblacksims #blacklivesmatter #blacksimsmatter
#kattwilliams
Technical Challenges

- Massive Size and High Velocity
- Highly Imbalance Data
- Ambiguous and Disguised Intent
- Multilingualism (script and language)
- Noisy Content (spell and grammar errors, abbreviations, slangs)
- Use of Multi-media formats
- Spam and Fake Accounts
- Data Annotation and Ground Truth
- Manipulation, Fabrication and Adversarial Behavior
"Yes I enjoyed and actually love CACW but I am so pissed off that the people who suffer / die in the end are a woman (Peggy dies) and two black men (T’challa’s father dies and Rhodes is paralyzed) Thank you marvel.”
Research Motivation and Aim

- **Research Motivation**
  - Performance of existing keyword spotting techniques
    - Presence of similar keywords
    - Presence of ambiguity - difficult for human annotation
    - Presence of hateful comments in a naïve post
  - Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

- **Research Aim**
  - To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
  - To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
  - To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Research Motivation and Aim

- **Research Motivation**
  - Performance of existing keyword spotting techniques
    - Presence of similar keywords
    - Presence of ambiguity- difficult for human annotation
    - Presence of hateful comments in a naive post
  - Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

- **Research Aim**
  - To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
  - To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
  - To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Research Motivation and Aim

**Research Motivation**
- Performance of existing keyword spotting techniques
  - Presence of similar keywords
  - Presence of ambiguity- difficult for human annotation
  - Presence of hateful comments in a naive post
- Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

**Research Aim**
- To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
- To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
- To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Research Motivation and Aim

Research Motivation

- Performance of existing keyword spotting techniques
  - Presence of similar keywords
  - Presence of ambiguity - difficult for human annotation
  - Presence of hateful comments in a naive post
- Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

Research Aim

- To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
- To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
- To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Research Motivation and Aim

Research Motivation

- Performance of existing keyword spotting techniques
  - Presence of similar keywords
  - Presence of ambiguity- difficult for human annotation
  - Presence of hateful comments in a naive post
- Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

Research Aim

- To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
- To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
- To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Research Motivation and Aim

- **Research Motivation**
  - Performance of existing keyword spotting techniques
    - Presence of similar keywords
    - Presence of ambiguity- difficult for human annotation
    - Presence of hateful comments in a naive post
  - Need to develop a solution for automatically identification and disambiguation of racist/radicalized intent posts

- **Research Aim**
  - To investigate the efficacy of NLP techniques on microblogging dataset for topic and intent classification
  - To investigate the application of various linguistic features for classifying the racist/radicalized intent posts
  - To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach
Prentice et al. [1] conduct a quantitative text analysis on 50 documents originated from extremist websites to study the persuasion behavior of online extremist media varying for the documents posted before and after the Israeli activities in Gaza.

Smith et al. [2] conduct a quantitative content analysis on public documents to distinguish radical groups from non-radical groups.
Given a dataset $D$ of Tumblr Posts $P_i$, $D = \{P_i \mid 1 \leq i \leq n\}$, a set of topics $N = \{N_j \mid 1 \leq j \leq m\}$ and a target class $C$; identify the intent of $P_i \in D$ when $P_i \in N$. We define a Tumblr post $P_i$ as a racist intent post if

1. the topic of the content belongs to a race or a religion and
2. the post targets a community in an offensive or persuasive manner (in a recognizable way)
Figure 1: Tag Search Based Bootstrapping Method For Extracting Tumblr Text Posts
Experimental Setup

Figure 2: Basic Statistics of English Language Posts from Experimental Dataset

Kappa Coefficient $\kappa$: Topic- 0.91, Intent- 0.95
Proposed Methodology

Features Identification

Sentiment & Semantic Enrichment using Open Source APIs

- Feature 1
- Feature 2
- Feature 3

Classification

One Class Cascaded Classifier

- Topic Classifier
  - Labeled Data, Target Class: Topic
  - Items labeled as Positive

- Intent Classifier
  - Labeled Data, Target Class: Intent
  - Intent Unknown

Supervised Classifier

Tumblr Text Posts
Proposed Methodology - Features Identification

- **Topic Modeling**
  - Language Taxonomy
  - Concepts and Terms

- **Sentiment and Tone Analysis**
  - Document Sentiment
  - Emotion Tone
  - Social Tone
  - Writing Tone

- **Semantic Tagging**
Proposed Methodology - Features Identification

- **Topic Modeling**
  - Language Taxonomy
  - Concepts and Terms

- **Sentiment and Tone Analysis**
  - Document Sentiment
  - Emotion Tone
  - Social Tone
  - Writing Tone

- **Semantic Tagging**
Proposed Methodology- Classification

- **Topic Classification**
  - religion and spirituality
  - society/unrest and war
  - society/racism
  - society/personal offense/hate crime
  - law, govt & politics/espionage and intelligence/terrorism and law
  - govt & politics/legal issues/human rights.

- **Intent Classification**
  - Document Sentiment
  - Emotion Tone
    (Joy, Sadness, Anger, Disgust, Fear)
  - Social Tendency
    (Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Range)
  - Language Cues
    (Analytical, Confident, Tentative)
  - Semantic Tagging
Proposed Methodology- Classification

- **Topic Classification**
  - religion and spirituality
  - society/unrest and war
  - society/racism
  - society/personal offense/hate crime
  - law, govt & politics/espionage and intelligence/terrorism and law
  - govt & politics/legal issues/human rights.

- **Intent Classification**
  - Document Sentiment
  - Emotion Tone
    (Joy, Sadness, Anger, Disgust, Fear)
  - Social Tendency
    (Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Range)
  - Language Cues
    (Analytical, Confident, Tentative)
  - Semantic Tagging
Proposed Methodology- Classification

- **Topic Classification**
  - religion and spirituality
  - society/unrest and war
  - society/racism
  - society/personal offense/hate crime
  - law, govt & politics/espionage and intelligence/terrorism and law
  - govt & politics/legal issues/human rights.

- **Intent Classification**
  - Document Sentiment
  - Emotion Tone
    (Joy, Sadness, Anger, Disgust, Fear)
  - Social Tendency
    (Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Range)
  - Language Cues
    (Analytical, Confident, Tentative)
  - Semantic Tagging
## Experimental Results

**Table 1:** Confusion Matrix for Topic Classification

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Topic</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>TP=253</td>
<td>FN=39</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>FP=93</td>
<td>TN=2034</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Performance Evaluation Metrics for Intent Classification.

<table>
<thead>
<tr>
<th></th>
<th>Test-Data1</th>
<th></th>
<th>Test-Data2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>RF</td>
<td>NB</td>
<td>DT</td>
<td>RF</td>
</tr>
<tr>
<td>Recall</td>
<td>0.79</td>
<td>0.82</td>
<td>0.79</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>Precision</td>
<td>0.72</td>
<td>0.78</td>
<td>0.74</td>
<td>0.75</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Experimental Results

**Table 1:** Confusion Matrix for Topic Classification

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Topic</td>
<td>Unknown</td>
</tr>
<tr>
<td>Topic</td>
<td>TP=253</td>
<td>FN=39</td>
</tr>
<tr>
<td>Unknown</td>
<td>FP=93</td>
<td>TN=2034</td>
</tr>
</tbody>
</table>

**Table 2:** Performance Evaluation Metrics for Intent Classification.

<table>
<thead>
<tr>
<th></th>
<th>Test-Data1</th>
<th></th>
<th>Test-Data2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>RF</td>
<td>NB</td>
<td>DT</td>
</tr>
<tr>
<td>Recall</td>
<td>0.79</td>
<td>0.82</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Precision</td>
<td>0.72</td>
<td>0.78</td>
<td>0.74</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Experimental Results

Figure 3: ROC Curve for Test-Data1 (Right) and Test-Data2 (Left)
Experimental Results- Relative Influence of Features

(a) Decision Tree

(b) Naive Bayes

(c) Random Forest

Figure 4: Percentage Fall in Accuracy of One-Class Classifiers During Leave-P-Out Compilation (P=1- One Feature)
Experimental Results - Relative Influence of Features

(a) Decision Tree

(b) Naive Bayes

(c) Random Forest

Figure 5: Percentage Fall in Accuracy of One-Class Classifiers During Leave-P-Out Compilation (P=2- Two Features)
Concluding Remarks

- We study the problem of identifying racist and radicalized Tumblr posts based on the intent of narrative.
- Results show that the proposed approach is effective for identifying intent-based radicalized posts in comparison to other keyword spotting methods.
- Authors' personality traits are a strong feature for identifying the intent of narrative.
- Topic classification of posts and filtering non-topic-based posts improves the accuracy of the proposed method.
Future Work

- Addressing the limitations of present study and improving the accuracy of linguistic features
- Sentence level identification of multilingual posts to improve the accuracy of translated content
- Identifying bloggers’ intent in multimedia posts (images and videos) by doing sentence detection in long tags
Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the gaza conflict.  

The language of violence: Distinguishing terrorist from nonterrorist groups by thematic content analysis.  
Thank You!
Contact Information: swatia@iiitd.ac.in
Web: www.swati-agarwal.in