Freedom of expressions provides leverage to an individual to share their opinions and beliefs about anything. However, many like-minded people misuse freedom of expression to make offensive comments or promote their beliefs that can lead to a negative impact on society [1]. Research shows that these individuals or groups of people use popular microblogging websites (Twitter and Tumblr) for such activities [2]. Existing literature shows that racism is not specific to only minor communities. There are users who post racist comments targeting existing like-minded groups calling it as reverse racism [3]. Based on our analysis, we broadly define these groups into two categories: Religion and Race. In this paper, we extend our previous work on extremist content detection on Tumblr microblogging website and address the challenge of mining intention of a narrative behind such posts [4]. We conduct a literature survey in the area of intent mining from free-form social-media text is technically challenging problem due to the presence of multilingual script, incorrect grammar, misspell words, short text, acronyms and abbreviations, sarcasm and opinion based posts. We also find examples of ambiguous and disguised content which makes a post difficult to classify even for human annotation. The work presented in this paper is motivated by the need to develop a system for automatically identifying a post made with racist or radicalized intent. In contrast to the existing work, our paper makes the following new contributions: 1) To the best of our knowledge, the study presented in this paper is the first work on racist and radicalization detection based on the intent of narrative unlike previous keyword spotting methods. 2) We apply natural language processing techniques on Tumblr posts for identifying discriminatory features for intent classification. 3) We publish the first ever semantically and sentimental enriched data of Tumblr posts and make our data publicly available for benchmarking and comparison [6].

## II. Experimental Setup

**Data Collection:** We conduct our experiments on an open source and real time dataset extracted from Tumblr microblogging website. We perform a manual inspection and create a lexicon of top \( K \) tags that are commonly used by racist or radicalized groups. For example, #islamophobia, #islam is evil, #supremacy, #blacklivesmatter. We implement a bootstrapping method to create our dataset and use this lexicon as seed tags for the Tumblr Search API\(^2\). For each tag, we extract only textual posts (text and quote) and extend our lexicon by acquiring other unique tags from extracted posts. We execute our model until we get a desired number of posts or the model converges (it starts extracting duplicate posts). Using Tumblr Search API, we were able to extract a total of 3,228 text posts made by 2,224 unique bloggers consisting of 10,217 unique tags. We remove all duplicate and non-English posts from our data and make it publicly available so that our experiments can be used for benchmarking and comparison [6].

**Data Annotation:** We use 2456 English language posts for annotation which spans only 83% of the extracted data.  

---

1. [http://dx.doi.org/10.17632/hd3b6v659v.2](http://dx.doi.org/10.17632/hd3b6v659v.2)
Since, we are using bootstrapping method to collect our data, it extracts a large number of noisy posts that do not belong to the defined topic (race and religion). Therefore, we first annotate the topic related posts and later label them as intent (racist/radicalized) or unknown. We employ two annotators with 2 to 3 years of experience of using Tumblr website and create ground truth for our data by measuring the inter-annotator agreement and Cohen’s Kappa coefficient between both annotations. Table I shows the results of topic and intent annotation performed on 2, 456 posts. Table I(a) reveals that we get 2, 419 (292 topic and 2, 127 unknown) posts as same label from both the annotators. We discard the remaining 37 posts with inconsistent annotation. Both the annotators further label these 292 topic posts as intent or unknown. Table I(b) reveals that the annotators agree on 278 posts (103 intent, 175 unknown) while there is an inconsistency in remaining 14 posts. Table I(c) shows the value of Cohen’s Kappa coefficient for both topic and intent annotation. Results reveal that the annotators agree more than 90% of the time. Annotation results shows that the intent posts are only 37% of topic posts and only 4% of the complete experimental dataset, revealing that the labeled data is highly imbalanced.

III. FEATURES IDENTIFICATION

Inspired by the prior literature, we create our feature space by analyzing the linguistic features of Tumblr posts and divide our features set into three categories: Topic Modeling, Tone Analysis and Semantic Tagging.

**Topic Modeling:** Existing literature shows that there has been a lot of work in the area of mining user generated content on social media related to offensive speech, racism and radicalization [4][7]. However, our analysis and annotation reveals that despite not having certain topic specific keywords, a post can be an intent post for which keyword based classification method do not work accurately and generates a large number of false alarms [2]. Therefore, we use statistical and natural language processing techniques to perform topic modeling on Tumblr posts. We use Alchemy Taxonomy API\(^3\) and Concept Tagging API\(^4\) to identify various concepts and classify the post into the most likely topic and sub-topic categories.

**Sentiment and Tone Analysis:** We investigate the language of a narrative by analyzing various types of sentiments and personality traits in a post. We conduct a linguistic analysis of text posts by using Alchemy Document Sentiment API\(^5\) and IBM Watson Tone Analyzer API\(^6\). Sentiment analysis identifies the positive and negative polarity of a post while tone analysis measures the level of three categories including emotion, social and writing tones. Emotions tone analyzes the text of a post and gives a distribution of 5 emotions namely joy, fear, sadness, anger and disgust. Social tendencies analyze the personality traits from the text including openness, conscientiousness, extraversion, agreeableness and emotional range of a narrative. Writing tone identifies the language cues of the author by measuring the analytical, confident and tentative style of writing.

**Semantic Tagging:** Semantic tagging of a post identifies the semantic role of each term present in the content. It also identifies the hidden phrases playing major role in the post. We use UCREL Semantic Analysis System (USAS)\(^7\) to tag each post in our dataset. All the semantic tags in a post are composed of a general or high level label and a numeric value showing the division of each label in lexicon. We remove all punctuations and special characters from semantically tagged content and decode all remaining terms with their respective labels in tags’ hierarchy.\(^8\) If a word is not identified by USAS, it is tagged as Z99. We however do not remove them from our dataset because USAS labels various topic specific terms as Z99 that are important for the intent identification. For example, ‘Jihadist’, ‘racial’, join words such as ‘BlackLivesStillMatter’.

IV. CLASSIFICATION

The third phase of our proposed framework is a cascaded ensemble learning based classifier primarily consisting of two stages: topic classification and intent classification. We train our model from feature vectors created in Section III and classify posts having racist or radicalized intent.

**Topic Classification:** We use topic modeling linguistic features of each post to identify the topics. We take a random sample of 50 posts from 292 posts annotated as topic and extract their taxonomy and concepts from the feature space. We create two independent lexicons of these concepts and labeled topics that has a confidence score above 0.40. We manually filter the list of taxonomy and finalize the following 6 labels that strictly belong to the topic of this study: 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. We use a look-up based method and check if the post belongs to any of these taxonomies and has a confidence score above 0.40. If yes, then we classify it as a topic post. If a post contains a wide range of taxonomies (>5) then we identify the top K concepts in the post and classify it as topic post based on their presence in the lexicon.

**Intent Classification:** An intent of a post cannot be fully determined only by mining the keywords in the content. But

---

\(^{3}\)http://www.alchemyapi.com/products/alchemylanguage/taxonomy

\(^{4}\)http://www.alchemyapi.com/api/concept-tagging

\(^{5}\)http://www.alchemyapi.com/api/sentiment-analysis

\(^{6}\)https://tone-analyzer-demo.mybluemix.net

\(^{7}\)http://ucrel.lancs.ac.uk/usas/

\(^{8}\)http://ucrel.lancs.ac.uk/usas/semtags.txt
it also requires to understand and predict the psychological tendency, sentiment tones and language of the narrative. It also requires to analyze the semantic role of topic related keywords used in the post. We perform classification on Tumblr posts by training our model on sentiment, semantic and language cues based features of a text. On a high level, we create a vector space of 5 features set (F1 to F5) which is further categorized into 15 unique vectors. Table II shows the list of all features extracted and grouped into 5 feature vectors. We define intent classification as a one-class classification problem. Therefore, our training data contains only positive class (intent) posts. We implement three different one-class classifiers (Random Forest (RF), Naive Bayes (NB) and Decision Tree (DT)) and compare their accuracy for the posts classified as topic in Stage 1. We train our model for each classifier and perform 5 fold cross validation. As discussed in Section II, only 12% of the posts are labeled as intent posts making our experimental dataset highly imbalanced. We select classification algorithms that works for small training data.

V. PERFORMANCE EVALUATION

In this Section, we present the accuracy results of each classifier discussed in Section IV and also present the influence of topic classification’s accuracy on intent post classification. Based on the inter-annotator agreement results, we evaluate the accuracy of our classifier by comparing the observed results against actual labeled class. We conduct our experiments on 2,419 posts and the proposed topic classifier classifies 346 posts as target (topic) class and 2,073 posts as unknown with a misclassification of 3.8% and 1.6% in identifying target and outliers (unknown) posts. Our results reveal that for topic classification, we are able to achieve a precision of 73% (253/(253+93)) and a recall of 86% (253/(253+39)). Given that our data is highly imbalanced and only 12% of the posts are labeled as target (intent) class, we execute each of our classifiers (RF, NB, and DT) using a 5 fold cross validation. Since, the accuracy measures are biased towards the majority class, we evaluate the performance of intent classifier using two standard information retrieval metrics i.e., precision and Area Under Operator Receiver Curve (AUC). Due to the misclassification in topic modeling, we evaluate the performance of intent classification in two steps. We execute our model on all 346 posts (Test-Data1) classified as topic in previous stage and 253 Tumblr posts (Test-Data2) correctly classified as topic. Table III shows the accuracy metrics for RF, DT and NB algorithms. Table III reveals that one-class intent classifier gives higher precision rate for Test-Data1 and filtering non-topic based posts from the dataset further improves the accuracy of intent classification. This is probably associated with the fact that unknown posts represent a broad range of sentiments and language cues. Table III reveals that Random Forest outperforms Naive Bayes and Decision Tree algorithms and gives the maximum precision (0.78, 0.81) and recall (0.82, 0.84) for Test-Data1 and Test-Data2. Our results show that wrongly classified posts at Stage 1 provokes a decrement in accuracy of intent classification. Figure 1 shows the ROC curves generated for each type of classifiers executed for both Test-Data1 and Test-Data2. Graphs in Figure 1 shows that given a set of posts \( P = \{P_i \mid 1 \leq i \leq n \mid C(P_i) = \text{Topic}\}\), Decision Tree based intent classifier has the high probability (~0.70) to classify them as target class. While, Random Forest and Naive Bayes have almost equal probability (0.55) to classify a post as intent or unknown. Figure 1 reveals that if the taxonomy of a post is unknown (Test-Data1) then each algorithm has a probability of approximately 0.60 to classify it as intent post.

In order to evaluate the impact of each feature on intent classification performance, we test the leave-p-out cross validation for both Test-Data1 and Test-Data2. Figure 2 illustrates the percentage of fall in precision of each classifier. Negative values rather shows the increment in precision. Figure 2(a) shows that in Test-Data1, removing F2 and F3 individually from the feature space does not impact the overall performance of Decision Tree (<1%). While removing writing tone feature (F4) decreases the precision by 4%. In fact, for Test-Data2, removing F1 from the feature space increases the performance of Decision Tree by 2%. It is possibly due to the reason because emotion tone gives a detailed classification of emotions while document sentiment feature gives overall sentiment of a post that can be biased in longer posts. Figure 2(b) reveals that in Naive Bayes algorithm, removal of any feature from Test-Data2 impacts the performance of classifier with a reasonably high percentage of fall in precision. Similarly, if the taxonomy of a post is unknown (Test-Data1) then removing F3 and F4 decreases the precision by 2%. Similar to Naive Bayes, for Random Forest algorithm (Figure 2(c)), removal of any feature declines the classifier’s performance up to 2%. While, for any

---

**TABLE II**

**FEATURE CODES AND GROUPING OF SIMILAR FEATURE VECTORS**

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Document Sentiment (1)</td>
</tr>
<tr>
<td>F2</td>
<td>Semantic Tagging (1)</td>
</tr>
<tr>
<td>F3</td>
<td>Emotions (5)</td>
</tr>
<tr>
<td>F4</td>
<td>Writing Tone (3)</td>
</tr>
<tr>
<td>F5</td>
<td>Social Tendencies (5)</td>
</tr>
</tbody>
</table>

**TABLE III**

**PERFORMANCE EVALUATION METRICS FOR INTENT CLASSIFICATION.**

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

---

**Fig. 1.** ROC Curve for Test-Data1 (Right) and Test-Data2 (Left)
unknown post, F3 and F4 are the most discriminatory features.

We also report the variation in performance of classifiers if a combination of two features is removed from the training model. Leaving out two features at once also reveals the relative influence of each vector in feature space. Figure 2(d) reveals that feature F3 and F4 are the most discriminatory features as removal of any of these vectors does not influence the performance of other features but we observe a fall in the overall precision rate. For example, removing feature F1 with F4 decreases the precision by 6% for Test-Data1 and 2.25% for Test-Data2. In Naive Bayes intent classification, we find that for Test-Data1, F2 is an important feature (Figure 2(e)). This is possible because if the taxonomy of a post is unknown then semantic tagging of text can be an important feature for identifying the topic related posts. Figure 2(d) also reveals that leaving out F3 and F4 features from training model individually makes a fall of 2% in overall performance while combining any of them with F5, the accuracy rather improves by 1% to 2%. It happens because if the posts are not topic related then they might have a wide range of taxonomy which impacts the social tone of a narrative. Due to the sparsity in social tendency attributes, it increases the number of false alarms. Unlike Decision Tree or Naive Bayes algorithms, in Random Forest, removing a combination of any two feature vectors decreases the performance rate of intent classifier for each dataset. Figure 2(f) reveals that removing any feature along with F3 declines the precision by at least 4%. While removing them with F4 can lower the performance by 2% to 4%. Our results reveal that emotion tone (F3) and writing cues (F4) are the two most discriminatory features for identifying intent post while using any of three classifiers and datasets. Semantic tagging (F2) and social tendency of narratives (F5) are two other important features if the post has a wide range of topics or emotional range making a post ambiguous.

VI. CONCLUSIONS AND FUTURE WORK

We formulate our problem of racist or radicalized intent classification as a cascaded ensemble learning problem and propose a two-stage one-class classification approach. Our result shows that the proposed methodology is effective for identifying intent posts unlike previous keyword based techniques. Our experimental results shows that Random Forest algorithm outperforms Decision Tree and Naive Bayes algorithms with a precision of approximately 80%. In this paper, we conclude that emotion tone, writing cues and social personality traits of an author are discriminatory features for identifying the intent of the post. Further, topic classification of posts and filtering non-topic based (or noisy) posts improves the performance of the proposed intent classification.

REFERENCES


