Investigating the Dynamics of Religious Conflicts by Mining Public Opinions on Social Media

Swati Agarwal[1] and Ashish Sureka[2]

[1] Indraprastha Institute of Information Technology, Delhi (IIITD), India,
swatia@iiitd.ac.in
[2] ABB Corporate Research Center, India,
ashish.sureka@in.abb.com

Abstract. The powerful emergence of religious faith and beliefs within political and social groups, now leading to discrimination and violence against other communities has become an important problem for the government and law enforcement agencies. In this paper, we address the challenges and gaps of offline surveys by mining the public opinions, sentiments and beliefs shared about various religions and communities. Due to the presence of descriptive posts, we conduct our experiments on Tumblr website- the second most popular microblogging service. Based on our survey among 3 different groups of 60 people, we define 11 dimensions of public opinion and beliefs that can identify the contrast of conflict in religious posts. We identify various linguistic features of Tumblr posts using topic modeling and linguistic inquiry and word count. We investigate the efficiency of dimensionality reduction techniques and semi-supervised classification methods for classifying the posts into various dimensions of conflicts. Our results reveal that linguistic features such as such as emotions, language variables, personality traits, social process, informal language are the discriminatory features for identifying the dynamics of conflict in religious posts.

Keywords: Mining User Generated Data, Public Opinions, Religious Conflicts, Social Computing, Text Classification, Tumblr, Semi-Supervised Learning

1 Introduction

Research shows that the with the unexpected emergence of religion and faith among people has also led to the discrimination and violence against rival religious groups [1]. It is seen that the people use various different platforms (chat groups, forums, blogs, social media) to share their beliefs and opinions about their religion [5]. These people also outburst their extremist and hateful views towards other religions [4][2]. These groups of individuals take the leverage of freedom of speech and social media to post their sentiments and beliefs about variety of sensitive topics including religion and race [2]. Despite several guidelines of social media platforms[3] and constraints of freedom of speech [8], people post racist and harsh comments against other religions that can hurt religious sentiments of an individual or community [4][10].

Figure 1 shows examples of several online posts showing the conflicts in

[3] https://www.tumblr.com/abuse/maliciousspeech

Fig. 1. A Concrete Example of 5 Tumblr Posts Showing Differences and Conflicts in beliefs of Tumblr Bloggers on Islam Religion

context of Islamic religious beliefs and sentiments of authors. These posts reveal that some users post defensive and promotional content about Islam religion. Whereas, some users posted negative comments and insulting the beliefs of people believing in Islam. Further, some users only make posts to share information on real time incidents or news and not presenting any sentiment or argument for a religion. As seen in the real world, many young age people and students get influenced from social media messages and join religious wars [5]. Therefore, monitoring such content on social media and identifying religious conflicts within society, understanding the root cause of such conflicts and arguments have become an important problem for the government, social scientist and law enforcement agencies.

Background: We conduct a literature survey in the area of political and religious conflict identification on social media. We find that over the past three decades, various social science researchers conduct offline surveys for identifying religious conflicts within society. Whereas, the area of identifying such conflicts by using computer science applications is not much explored. Based on our analysis, we divide our literature survey into four lines of research:

1. Offline Data and Manual Analysis: Swinyard et. al. [12] and Wilt et. al. [15] conducted surveys to examine the relation between religious and spiritual beliefs and emotions of people such as happiness and anxiety. Yang et. al. [16] present a study on the impact of low coverage of HindRAF event in media causing the religious conflicts among citizens of Malaysia.

2. Offline Data and Automated Analysis: Vüllers et. al. [14] present a study on the religious factors of 130 developing countries. Their analysis reveal that the clashes between religious groups and attacks by religious actors are the main cause of religious conflict within state and community. Basedau et. al. [6] used logistic regression approach on the same dataset to identify several discriminatory religious factors that causes conflicts, religious violence and grievances.

3. Online Data and Manual Analysis: In addition to social science researchers, various non-profit organizations like Pew Research Center[4], Berkley Center for Religion, Peace and World Affairs [5] and United States Institute of Peace[6] conduct online polls, offline statistical and text analysis on blogs and social media data to identify religious beliefs and issues within local and global

[4] http://www.pewresearch.org/topics/religion-and-society/
[5] https://berkleycenter.georgetown.edu
[6] http://www.usip.org/about-usip

regions. Some of the recent studies of Pew Research Center include the global trend and projection of population growth of various religions, gender gap in religious commitment of Muslim and Christian communities and increment and decrement rate of government restrictions on religion and social hostilities.

4. Online Data and Automated Analysis: Chesnevar et. al. [7] propose an opinion tree using information retrieval and argumentation technique for identifying conflicts and confronting opinions in E-Government contexts. They conduct their analysis on Twitter messages and identify the polarity (positive, negative and neutral) of contrasting arguments. Agarwal et. al. [4] conduct a manual analysis on Tumblr posts to investigate the feasibility of content analysis for identifying religious conflicts and fill the gaps of traditional offline surveys.

Motivation: The work presented in this paper is motivated by the prior literature and a need to develop an automatic solution to identify religious conflicts among social media users. However, automatic identification of religious beliefs and faith by mining user generated data is a technically challenging problem. In order to enhance our understanding of religious conflicts and address the challenge of local and regional data, we conduct our experiments on a wider community of Tumblr. Tumblr is the second most popular micro-blogging service that allows users to post eight different types of content including image, video, audio, chat, quote, answer, text and url [2]. Unlike Twitter, Tumblr has no character limit for tags, image captions, text body content allowing it's users to make descriptive posts. Presence of noisy content such as misspelt words, short text, acronyms, multi-lingual text and incorrect grammar decreases the accuracy of linguistic features and Natural Language Processing tools [5]. Further, the presence of ambiguity in posts and the intent of author makes it difficult even for human annotation [2]. We, however conduct our analysis on Tumblr website because Tumblr allows users to make longer posts and express their opinions and beliefs in an open and descriptive manner which fills the gaps of offline surveys. Furthermore, Tumblr facilitates it's users to send anonymous messages and use the leverage of expressing their opinions without revealing their names.

Research Contributions: In contrast to the existing work our paper makes the following novel and technical contributions: 1) To the best of our knowledge, we present the first study on automated identification of religious beliefs, opinions and faith in global public communities. 2) We address the challenge of social media content by translating multi-lingual posts into base language and extracting textual metadata of multimedia posts such as photo and video. 3) We identify various linguistic features that are discriminatory for identifying contrast in different opinions on religious posts and 4) We investigate the efficiency of multi-class semi-supervised classifier across various dimensionality reduction techniques for classifying Tumblr posts into various dimensions of conflicts.

2 Dimensions of Conflicts

We conducted a survey among 3 different groups of people- we selected 10 graduate students of our department, 30 Tumblr bloggers (followers on authors' personal Tumblr account) and 20 people from society randomly. Following the various facets presented in Agarwal et. al. [4], we conducted a small questionnaire

Table 1. Concrete Examples of 11 Dimensions and 3 Polarities of Religious Beliefs and Sentiments in Tumblr Posts Created About Christian Religion and Community

Type	Post Content
IS	In a show of solidarity, Muslims are standing with Christians and giving up guilty pleasures for lent.
Query	Doesn't the Bible teach us not to take a life of another? To turn the other cheek and not respond with violence? Isn't better to die and be in heaven then kill and stay on earth?
N/A	Pray for abortion access. People deserve easy access to abortion services.
Defensive	I'm still over the moon about God. I'm in total awe that He not only hears me, but actually listens and does something about it. I feel so loved and acknowledged.
Disappointment	If you're a Christian and voted for Trump I wanna ask you a question. What does it feel like to go against everything God wanted for us?
Annoyance	Jesus himself could crawl out of his grave, take me by the hand, and point me to salvation and heaven. I would say no. I would seriously 100% rather die as a Jew then live for even a millisecond as a Christian. So stop trying to convert me to Christianity because it is not going to happen.
Insult	Burn churches not calories. Christianity is stupid!- Well I am not the only one that feels the same way.
Disgust	So this dude that was running in local elections for council said women who have abortions are worse than ISIS
Ashamed	I feel like a bad Christian. I have so much hate in my heart after this election, at Drumpf, at his voters, at my country. I know I should turn the other cheek and love radically and protest without hating but I'm so angry. I feel like I can't let that hate go, not so soon. But I need to and I'm furious at myself
Disbelief	Imagine the peace we'd all have without religion. Wouldn't it be a better world?
Sarcasm	When Christ has a cold he sneesus

consisting of questions related to their activities on social media platforms. How frequently they make religion based posts on social media or react to other religious posts. We created a set of 30 posts about different religions and asked for their opinions on these posts. Based on the dimensions discussed in Agarwal et. al. [4] and our survey, we decide 11 dimensions of opinions that can be used to define the contrast of conflict among people: Information Sharing (IS), Query, Not a religion based post (N/A), Disbelief, Defensive, Annoyance, Insult, Disappointment, Sarcasm, Ashamed and Disgust. Table 1 shows the examples of 11 Tumblr posts created about Christian religion and community reflecting the different dimensions of public opinions about the community.

3 Experimental Setup

Acquiring the Dataset: To conduct our experiments, we download a publicly available dataset [3] published by Agarwal et. al. [4] on Mendeley[7]. As of November 9, 2016, this dataset is the largest dataset available of Tumblr posts and bloggers. The dataset contains all types of Tumblr posts (answer, photo, text, audio, video, url, chat and quote) consisting of various tags frequently used in religion based posts. The published dataset contains a total of 107,586 posts collected for 10 such tags (hinduism, islam, muslim, religion, isis, jihad, christian, islamophobia, judaism and jews). The statistics reveal that the maximum number of posts consisting of religious tags are either posted as photo (49,072) or text (34,902) posts. Similarly, URL or link types of posts (10,062) are relatively higher in comparison to chat (507), audio (390) and answer/ask box (1,077) categories [4].

Data Pre-processing: In order to identify the religious conflicts, we conduct our analysis only on textual metadata of posts. Therefore, in this phase, we address the challenge of multi-lingual and multi-media content of the posts. In

[7] `http://dx.doi.org/10.17632/8hp39rknns.1`

Tumblr, each type of post contains a different set of textual attributes. We acquire different metadata of all records acquired according to the type of the post available in our experimental dataset. For example, for photo and video posts, we extract only the caption and description of posts, for chat and answer posts, we extract the phrases used in the conversation. Similarly, for URL, quote and text posts, we extract the title and body content of the posts. We discard the audio posts since these post contains only track name, artist name and album name which do not reveal any information about the post. In this paper, we conduct our experiments only on English language posts. Therefore, in order to address the challenge of multi-lingual posts, we translate all non-English posts into our base language. We use Yandex Language API[8] to detect the language of source content and translate it to English language. We further remove the posts consisting of no textual metadata. For example, photo posts with no caption or text posts consisting of only external URLs. We also remove all redundant posts from the dataset consisting of different post_id but having duplicate content. After pre-processing of the raw data, we were able to acquire a total of $89,803$ posts calling them as our experimental dataset.

Data Annotation: In order to create the ground truth for our dataset and creating a training dataset, we use 89,803 pre-processed posts for further annotation which spans only 83.4% of the acquired data. In order to remove the bias from our annotation, we hired a group of Tumblr users who had an experience of 2-3 years of using Tumblr website. We published a post on Tumblr and asked bloggers to volunteer for data annotation. In a span of one week, 34 bloggers replied and agreed to annotate an average of 30 posts. We declined 4 bloggers who joined Tumblr recently. Among 30 bloggers, only 23 bloggers reverted back with 690 annotated posts among which 6 posts were sampled more than once. Due to the large amount of Tumblr posts and challenge of creating ground truth [5]; we used only these 684 posts for creating our training dataset.

4 Features Identification

Topic Modeling: During our survey for identifying the dimensions of conflict, we observe that many users add religion based tags in their posts while the content of the post is irrelevant to any religion or community. Since, our experimental dataset is collected using a keyword based flagging approach, we identify the topic of each post to filter the irrelevant posts. Figure 2 shows the statistics of number of posts consisting of religion based tags and actually discussing about those religions. Figure 2 reveals that among all posts (85% of experimental dataset) consisting of seed tags related to Islam religion (islam, muslim, islamophobia, isis and jihad), only 31% of the posts are about Islam religion. Similarly, among $20,106$ posts (22% of experimental dataset) consisting of judaism and jews tags, only $15\%(13,695)$ posts belong to Judaism religion. We implement the algorithm proposed in Agarwal et. al. [2] for identifying the topic of Tumblr posts. For each post, we assign a binary value where 1 denotes

[8] `https://translate.yandex.com/developers`

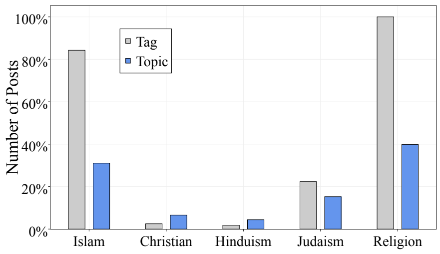


Fig. 2. Relative Percentage of Number of Posts Consisting of a Religion Based Tag and Topic

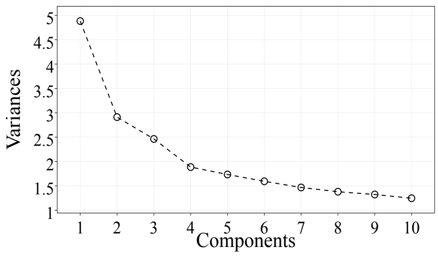


Fig. 3. Relationship Between Variance and Components in Principal Component Analysis

the topic (religion based post) and 0 denotes the non-topic (not a religion specific post). We further extract the name of the religion being discussed in the post since a post can have content about more than one religion. For example, in the following post *"KKK burns black Churches even tho they claim Christianity as their religion and ISIS blows up mosques even tho they claim Islam as their religion."* author mention about both Islam and Christian religion. In our experimental dataset, we find that only 40% of posts (35,799) belongs to a religious topic (Islam, Hinduism, Christian and Judaism) while the remaining 60% of posts (54,004) only contains religious tags but do not contain the content related to a religious group or community.

LIWC: In order to compute the correlation between various linguistic features and sentiments, we use an open source API by LIWC- Linguistic Inquiry and Word Count [13]. We extract a total of 45 features grouped into 14 categories of linguistic dimensions. In order to identify the sentiments and emotions of bloggers, we compute the relative percentage of emotions i.e. sadness, anxiety, anger, happiness. We compute the authenticity and personality traits of authors by computing summary of language variables (analytical thinking, authenticity) in a post. Further, in order to identify the personal beliefs and relation with the real world incidents, we compute the percentage of sexual terms, mention of family, friends, male and female references in a post. In order to identify the level of aggression and certainty of a post, we compute the percentage of use of informal language such as the presence of swear words, slangs and fillers. Apart from these features, we also compute the presence of various other linguistic dimensions such as pronouns, negations, interrogatives words, cognitive process, perceptual process, power, time orientations (mention of past, present or future incidents), time and personal concerns (work, religion, death).

5 Features Selection

1. Using All Features (FS1): In first iteration, we use all 45 linguistic, sentiment and text based features extracted using LIWC. We train our model on available features and investigate the efficacy of classification of Tumblr posts into 11 dimensions of conflicts.

2. Principal Component Analysis (FS2): In second iteration, we use

PCA- a dimensionality reduction technique to reduce the number of feature vectors. In PCA, we compute the correlation among all features vectors and identify the components chracterizing the whole data. For $n = 45$ feature vectors in our experimental data, we get n eigenvectors. We select the first $p = 10$ eigenvectors having maximum eigenvalues and discard the eigenvectors with less significance. Therefore, we project our dataset into 10 dimensions. We form our feature vector $FS2$ by taking the eigenvectors of 10 components. Figure 3 shows the distribution of variances for all 10 components selected after dimensionality reduction of the data.

3. Attribute Selection Correlation(FS3): In third iteration, we use Correlation Attribute Evaluation technique to identify a set of discriminatory attributes. We measure the Pearson's correlation between each attribute (feature vector) and the class. We create a correlation matrix of 45 attributes and class for each record in the dataset and compute the overall correlation by computing the weighted average of the attribute. Based on the correlation between each attribute vector and class, we create a set of 10 features having moderately higher positive and negative correlation and drop the features having correlation closer to zero. For our experimental dataset, FS3 returns the following 10 features: mention of past and present tense, pronouns, male references, perceptual process, negative emotions, clout, presence of negation, swear words and anger.

6 Classification

Classes and Membership Groups: Based on the polarity of opinions in religious posts and their importance in defining the dynamic of conflicts, we split the dimensions of conflicts into six classes: information sharing, query, N/A, sarcasm, defensive and disagreement. We further divide disagreement class into six subclasses reflecting a higher range of negative emotions: disappointment, annoyance, insult, disgust, ashamed and disbelief. For a given data point y_m, in order to identify the the polarity of the post, we first classify the post into six classes and assign a label o_m. If the post is identified as a disagreement or negative post, we further classify into six subclasses identifying the low-level details of negative emotions in a given post.

Classification Approach: Due the constraint of lack of ground truth and only a very small portion of available labeled data (2%), we use semi-supervised classification method to classify the unlabelled posts over unsupervised method. Semi-supervised classification approach uses both annotated and unlabelled data to learn the model iteratively in a snowball manner. We use 684 posts annotated by Tumblr bloggers and use them to train our model in first iteration of semi-supervised classifier. We conduct our experiments on $35,799$ posts identified as topic related (discussing about any religious group or community). Given a labeled data (X_N, C_N) where the data points are denoted by $X_N = (x_1, x_2, x_3...x_n)$ and their labels are denoted by $C_N = (c_1, c_2, c_3...c_n)$. The unlabelled data points $Y_M = (y_1, y_2, y_3...y_m)$ and their unknown labels $O_M = (o_1, o_2, o_3...o_m)$ are denoted as (Y_M, O_M).

We use the *'R'* statistical language to perform classification using *"upclass"* package [11]. "Upclass"[9] is a semi-supervised classification method and an adaptive version of the model-based classification method proposed in Dean et. al. [9]. Upclass uses an iterative method that initiates by using model-based classification method and uses Expectation- Maximization (EM) algorithm in further iteration until convergences. In the first iteration of classification, a set of 14 models is applied on the dataset considering different constraints (E- equal, V- variable, I- identity) upon covariance structure- volume, shape and orientation of the cluster. For example, in EEE model each cluster has equal volume, same shape and same orientation along the axis. The clustering is performed in multiple iteration by estimating group membership on unlabelled data based on the maximum likelihood of EM algorithm. In order to perform the model based discriminatory analysis on unlabelled data points, the model of data (combination of E, I, V constraints) must be known. If the model is null then Upclass fit every model to the data and identifies the best-fitted model of given data and attributes. To identify the best-fitted model, Upclass calculates the bayesian information criterion (BIC) value for each model. $BIC = 2\log(l) - p\log(n)$; where l is the likelihood of the data, p is the number of model parameters and n is the number of data points. The model with the highest BIC value is selected as best-fitted model for the data.

7 Empirical Analysis and Evaluation Results

In this Section, we present the classification results of Upclass semi-supervised method applied for both classes and sub-classes identification. We apply 3 iterations of Upclass supervised classification methods on all $35,799$ posts for each feature vectors model (FS1, FS2 and FS3) discussed in Section 5. If a post is labeled as "Disagreement or Negative", we further train our model on the posts labeled under the six subclasses of disagreement and classify unknown data points into one of the six sub-groups using Upclass semi-supervised classification method. Table 2 shows the experimental results of classification performed using each feature vector model for each membership groups. Table 2 reveals that the classification model converges for each set of feature vectors. During the first step of classification both FS1 and FS3 takes similar number of iterations whereas, FS2 takes approximately 2.5 times of their iterations. Further, for FS1 and FS3, VEV is selected as the best-fitted model while for FS2 attributes, VVV showing the non-linear distribution of labels (different orientation of each cluster against the axis). Figure 4 shows the visual representation of clusters created using different models (considering the constraints on covariance structure). Table 2 also reveals that during the second set of classification (sub-groups of disagreement class), Upclass method takes a different number of iteration for each feature vector model. Further, for each feature vector, a different discriminant model is selected. While, in FS2, for both classes and sub-classes identification Upclass uses the same model i.e. VVV. While using all attributes as features, it classifies all observations into equal parts and create clusters of equal shapes varying in

[9] `http://CRAN.R-project.org/package=upclass`

Table 2. Classification Results All Feature Selection Techniques for Different Membership Groups and Observations

	FS1		**FS2**		**FS3**	
Attribute	**Class**	**Sub-Class**	**Class**	**Sub-Class**	**Class**	**Sub-Class**
Converged	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Iteration	272	491	604	252	207	110
Dimension	45	45	10	10	10	10
Model Name	VEV	EEV	VVV	VVV	VEV	VVI

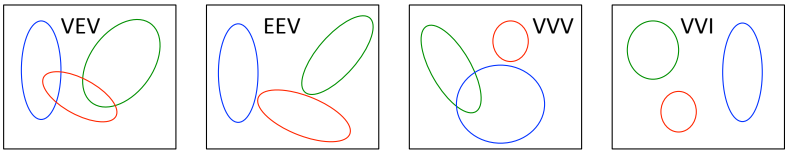
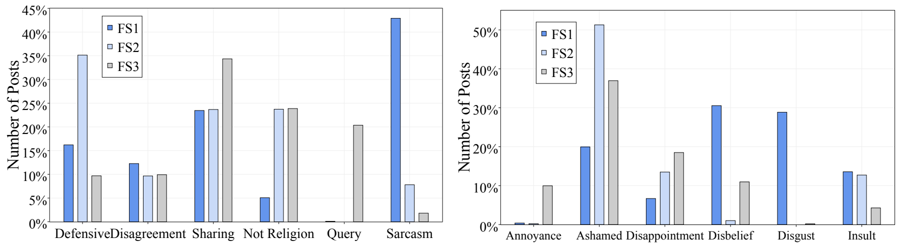


Fig. 4. Visualization of Volume, Shape and Orientation Constraints for Best-Fitted Models for Dataset Classification. V= Variation, E= Equal and I=Identical



(a) Distribution of Polarity Based Classes Identified for All Observation Posts in Experiemental Dataset

(b) Distribution of Sub-Classes Identified for All Posts Labelled as "Disagreement/Negative"

Fig. 5. Classification Results of Upclass Semi-Supervised Method for Unknown Posts Categorized into Polarity Based Classes and Extreme Emotions Based Sub-Classes

the orientation against axis. While, using features selected using Pearson's correlation technique, it classifies the observations in a linear manner- varying the shape and volume of the clusters while all data points aligned towards an axis. Unlike, FS1 and FS3, while using principal components as feature vectors the clusters are created in a non-linear manner- varying in size, shape and orientation of data points.

Figure 5 shows the distribution of Tumblr posts classified into different groups of classes and sub-classes based on the polarity of opinions. Figure 5(a) shows the relative percentage of posts classified into each of the defined classes. Figure 5(a) reveals that while using all attributes as features, maximum number of posts (43%) are labeled as sarcasm posts which is below 10% while using dimensionality reduction techniques. While only a very small percentage of posts (5%) are classified as non-religion based posts which is significantly higher for both FS2 and FS3 (approximately 25%). The classification results shows that except FS3, using FS1 and FS2 feature vectors, the classifier does not have sufficient examples for labelling query posts. The graph in Figure 5(a) shows that

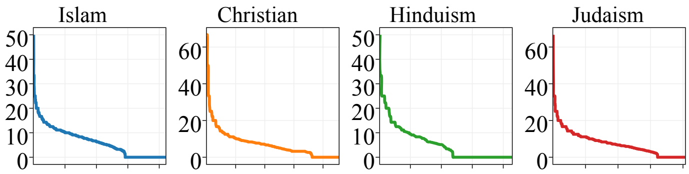


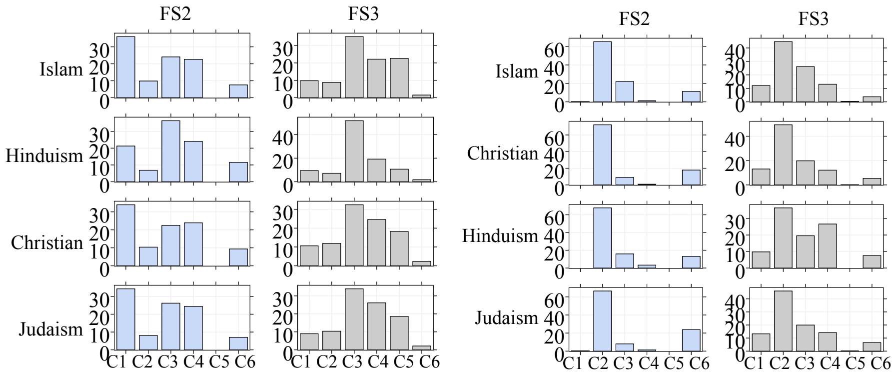
Fig. 6. Distribution of Negative Emotions for All Posts Identified as Disagreement or Negative Using PCA Feature Vector Model

for each feature selection method, the classifier classifies 10% to 12% posts as disagreement/negative that are further classified into sub-classes. Figure 5(b) shows the relative percentage of these 10% to 12% posts further classified into sub-classes of extreme negative emotions. Figure 5(b) reveals that while taking all attributes into account, a very small percentage ($\sim$ negligible) of posts are classified as "Annoyance" posts while the distribution of other classes are significantly higher. While the distribution of posts for FS2 and FS3 is varying for each category- as reflected in best-fitted model selected for classification (refer to Table 2. The variation in distribution of all posts in different categories shows the dynamics of public opinions on religious posts. The size of each cluster (number of posts grouped in a class) for different combinations of attribute selection techniques and classification method shows the presence of religious conflicts among users on Tumblr.

Figure 6 shows the distribution of negative emotions computed for all posts identified as disagreement and discussing about different religions. X-axis shows the number of posts while Y-axis shows the computation score of negative emotions in a post. Figure 6 reveals that for Islam among 2000 posts identified as disagreement, a large percentage of posts has higher rate of negative emotions (reaching upto 50%) which is even lesser than the negative emotions used in Christian and Judaism religions (reaching upto 60%) based posts reflecting the higher rate of hatred and discrimination among users.

In order to address the challenge of identifying public beliefs and opinions in Tumblr posts where authors are discussing about more than religion. We identify the name of religions being discussed in each post available in our experimental dataset. We classify each post into classes (polarity based groups) and sub-classes (extreme negative emotions based groups) and discuss the results of classification for identifying religion specific conflicts among Tumblr users. Due to the large volume size of Sarcasm cluster and no post classified as Query post, we discard the FS1 technique for identifying the conflicts among individual religious groups. Figure 7 shows the classification results and distribution of Tumblr posts classified into various dimensions of conflicts. For Figure 7(a), C1, C2, C3, C4, C5 and C6 denote defensive, disagreement, sharing, not religion, query and sarcasm dimensions respectively. Similarly, for Figure 7(b), C1, C2, C3, C4, C5 and C6 denote annoyance, ashamed, disappointment, disbelief, disgust and insult.

As shown in Table 2, while using principal component analysis feature vectors, the semi-supervised classification method selects VVV as best-fitted model. Figure 7(a) also reveals that for FS2 feature vectors the volume of all clus-



(a) Distribution of Polarity Based Classification of Tumblr Posts

(b) Distribution of Sub-Classification Results of Tumblr Posts

Fig. 7. Distribution of Classification Results of Tumblr Posts Specific to A Religion.

ters are different making some cluster too large or too small. Further, Figure 7(b) reveals that maximum number (more than 60%) of disagreement posts belong to "ashamed" category. Whereas, while using Pearson's correlation selection method, the posts are grouped into all classes. During the first phase of classification, there is variation in volume of observations in each cluster and while the shape of each cluster is the same. Whereas, in second phase of classification, the volume of observation and shape of each cluster varies. The classification of Tumblr posts for each religion into several dimensions of conflict shows that a lot of discussion about religious topics happen on social media where users have different opinions, beliefs and sentiments about these religions. Our results shows that various linguistic features, emotions, social presence, summary of language variables and other linguistic dimensions of user generated data can be used to identify the conflicts within religious faith and beliefs. Furthermore, Tumblr is a rich source of collecting public opinion posted in a descriptive and open manner which is useful to study the low-level details of religious beliefs and overcome the challenges of offline data and surveys.

8 Conclusions and Future Work

Research shows that due to the rapidly growing influence of religious faith and beliefs and leading to discrimination and violence against other rivalry communities, identification of dynamics of religious conflict has become an important and challenging problem for the government and law enforcement agencies. In this paper, we address the challenge of offline surveys by mining the public opinions and beliefs from Tumblr website. We conduct our experiments on an open source dataset consisting of the largest collection of Tumblr posts. We conduct a survey among three different groups of people (graduate students, Tumblr bloggers and people from society) and define 11 dimensions of public opinion that can identify the contrast of conflicts. We investigate the feasibility and efficiency of linguistic features and different dimensionality reduction techniques and compare their results of classifying Tumblr posts into different dimensions of conflicts. Due to the small size of labelled data, we use Upclass- a semi-supervised classification method to train our model and classify unlabelled observations. Based on our results, we conclude that despite the presence of noise and ambiguity in content, linguistic features are discriminatory features for identifying the dynamics of religious conflicts. Furthermore, identifying the topic prior to the identification linguistic features can be used to disambiguate the sentiments of author while discussing about more than one religion in a single post. Future work includes the improvement in linguistic features and making them efficient for classifying very short and short text posts. Furthermore, future work includes the identification of age and location of bloggers for identifying the collision of religious beliefs and sentiments in different age groups or different regions across the world.

References

1. Acciaioli, G.: Grounds of conflict, idioms of harmony: custom, religion, and nationalism in violence avoidance at the lindu plain, central sulawesi. Indonesia (72), 81–114 (2001)
2. Agarwal, S., Sureka, A.: But i did not mean it!- intent classification of racist posts on tumblr. In: Proceedings of European Intelligence and Security Informatics Conference (EISIC). IEEE (2016)
3. Agarwal, S., Sureka, A.: Religious beliefs on social media: Large dataset of tumblr posts and bloggers consisting of religion based tags, mendeley data, v1, http://dx.doi.org/10.17632/8hp39rknns.1 (2016)
4. Agarwal, S., Sureka, A.: A collision of beliefs: Investigating linguistic features for religious conflicts identification on tumblr. In: Proceedings of 13th International Conference on Distributed Computing and Internet Technology, ICDCIT, Odisha, India. Springer (2017)
5. Agarwal, S., Sureka, A., Goyal, V.: Open source social media analytics for intelligence and security informatics applications. In: Big Data Analytics (2015)
6. Basedau, M.: Bad religion? religion, collective action, and the onset of armed conflict in developing countries. Journal of Conflict Resolution 60(2), 226–255 (2016)
7. Chesnevar, C.I., Maguitman, A.G.: Opinion aggregation and conflict resolution in e-government platforms: Contrasting social media information. Interdisciplinary Perspectives on Contemporary Conflict Resolution p. 183 (2016)
8. Cohen, J.: Freedom of expression. Philosophy & Public Affairs 22(3) (1993)
9. Dean, N., Murphy, T.B., Downey, G.: Using unlabelled data to update classification rules with applications in food authenticity studies. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55(1), 1–14 (2006)
10. Emmons, R.A.: Assessing spirituality through personal goals: Implications for research on religion and subjective well-being. Social indicators research (1998)
11. Russell, N., Cribbin, L., Murphy, T.B.: upclass: An r package for updating model-based classification rules (2012)
12. Swinyard, W.R., Kau, A.K., Phua, H.Y.: Happiness, materialism, and religious experience in the us and singapore. Journal of happiness studies 2(1), 13–32 (2001)
13. Tausczik, Y.R.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)
14. Vüllers, J., Pfeiffer, B.: Measuring the ambivalence of religion: Introducing the religion and conflict in developing countries (rcdc) dataset. International Interactions 41(5), 857–881 (2015)

15. Wilt, J.A., Grubbs, J.B., et. al: Anxiety predicts increases in struggles with religious/spiritual doubt over two weeks, one month, and one year. The International Journal for the Psychology of Religion (2016)
16. Yang, L.F., Ishak, M.S.A.: Framing interethnic conflict in malaysia: A comparative analysis of newspapers coverage on the hindu rights action force (hindraf). International Journal of Communication 6, 24 (2012)