The authors would like to start by thanking the reviewers for their valuable input. We have integrated most of the comments into the manuscript. Below are the reviewers' comments and our responses indicating what we have done (in blue). Below are the main reviewers’ suggestions and actionable comments, followed by our comments on the changes we integrated to the paper.
Thank you again.

**Reviewer 1 main actionable comments:**

Removing few number of features from the a very large feature space would typically not have much impact on the performance unless these features have high discrimination power in comparison with all other features, which has not been measured or clarified in this work.

*As requested by the reviewer, we removed the tweets containing on-topic words (islam and muslim) along with the related network features for users who had an opinion before the event. The performance for content features didn’t significantly changed but for network-based features it dropped by about 40% which indicates the importance of network features within on-topic tweets. We added the text beginning with: “We also repeated the experiment for US users who expressed their opinion … “*

Secondly, according to the behaviour consistency hypothesis that the authors rely on, a user’s stance or behaviour is unlikely going to change over a small window of time, which in turns implies that the user’s language, profile and network features/characteristics are not going to substantially change either. As a result the classifier’s performance itself is unlikely to be affected. - Pertaining to the above point, the authors used two datasets to measure and compare the classifier’s performance with the presence/absence of topic-related words. The first dataset contains topics-related words while the other doesn’t. Such evaluation and comparison in my in opinion is not very sound due to the different size and characteristics of these datasets. I suggest the authors repeat the evaluation using a single dataset by keeping/excluding the topic-related words with each run.

*Addressed in section 5.4, by removing topic-related tweets for users who talked on-topic which resulted in significant drop in performance for network-based features.*

Also, believe it would be very useful to measure and report the discrimination power of these words for stance classification using IG or other proper measurements.
We added figures 6-13 at the end of the paper which list the most discriminating terms, based on the SVM scores, for each of the feature types. We put them in the appendix as not to break the flow of the paper.

- The stance class distribution in the dataset used for evaluation is rather imbalanced with 77% users of positive stance and 23% of negative stance. Consequently, the evaluation results were biased toward the positive class. For better evaluation and cross-comparison of the features types I suggest the authors to balance the number of positive and negative samples in the dataset and repeat the experiments.

Because we evaluate the method over three countries each with two sets of users (users who spoke on topic or not before the event), we evaluated the sentiment prediction method over each of the 6 datasets using AUROC so that the results can be compared over all the datasets.

We used the AUROC evaluation measure in all experiments so the results can be cross-compared. AUROC added to all the result tables.

This would lower the impact of the imbalance class distribution on the classifier’s performance as well as give more accurate evaluation. - The number of tweets used for training is rather small (200 tweets per user) in comparison with the number of samples (users). This may produce a sparse feature space for the three types of features and affect the classification performance. Please report the number of features used for training under each feature type. Also, I suggest to train the classifiers from the tweets posted after the event (177K tweets), as this would (i) give an idea of stance detection performance before and after the event, and (ii) allow for better evaluation and comparison between the three types of features.

The number of features are as follows: network: 15467 profile: 1749 content: 50418 (for US users who talked on-topic before the event). We add the text in the paper beginning with: “The content has the largest number of features followed …”

- Have you tested the statistical significance of your results? This should be reported in order to conclude a statistically significant improvement of the proposed features upon baselines.
We added the following to the paper:

We also evaluated the statistical significance of the results for each feature type using random permutations (Ojala et al., 2010) and found all the models to be significant at 0.01 level except for the models that only use profile field features which is not a surprise given that there is not enough signal about the sentiment of the users in profile fields (name, location and description).

We evaluated the significance of the results for each feature type using random permutations (Ojala, 2010) and found that except profile features, other feature types are significant at 0.01 level (section 5.4).

Data Collection and Annotation- Section 3.2 It’s unclear whether the tweets in the manually annotated sample are unique or contains any duplicates?

Manual annotation was applied to only unique tweets in samples. This was explained in the text. Table 1 was updated with the numbers of unique tweets in each sample.

Also, were all these tweets geo-located or not? Please clarify. - It is unclear how the location identification algorithm proposed works. As described in Section 3.4.1, the algorithm runs on all the 336,294 tweets to detect their location and those tweets of that the algorithm failed to detect their location were manually geo-located. If this is the case, why do you need to run the algorithm again as described in Step 5 in page 5?

Regarding the geo-location, this was applied on the user level not the tweet level. Actually the same tweet can be retweeted from different users in different locations. This seems to be not clear in our text. We have updated it to make it more clear. Now, it is clear that step 5 is only applicable for users with declared location. For users with blank or undefined location, the tweet content-based algorithm in section 3.4.2 is applied. We also added text here to declare it more.

- To extend the set of 107K geo-located tweets identified using the proposed algorithm, the authors identified the location of additional 70K tweets using a third-party text-based method. However, this method is trained from English tweets of different domain and topic, and hence it does not seem very sound to apply this method on the multilingual dataset used in this work. The accuracy of this method is 77% on the English data which is trained from as described in Section 3.4.2. This means that the accuracy is likely drop when the method is applied to tweets domain and multiple languages. I
suggest the authors to measure the accuracy of this method using a sample of their dataset.

*We added numbers for country-level for the top 10 countries and reported geolocation accuracy. We added Table 2 and the text beginning with: “We evaluated the performance of the text-based geolocation by comparing …”*

-I find the ranking of countries based on the users’ stances (Section 4.2) kind of biased and misleading due to the imbalanced distribution of tweets across the reported countries. E.g., the US has 36.5% of the tweets while a country like Netherlands has less than 1%. As such, It is not very sound to conclude that Netherland shows a negative stance against Islam while the US shows a positive one, as the data sample used for ranking is skewed and not representative of the countries’ positive and negative attitudes. I suggest you remove this section or rewrite by excluding the strong statements and conclusion about the reported rankings.

*We added a statement to make it weaker (especially for the countries under Singapore)*  
*We calculated the confidence interval for estimated percentage of support for all the countries in figure 5 (using the following tool: http://www.surveysystem.com/sscalc.htm), and we added the following paragraph for declaration:*

> “We calculated the confidence interval for each of the countries when setting the confidence level to 95%, since a sample of 100 tweets only is considered low to represent a country of populations in millions. It was found that most of the countries has a confidence interval of less than 5%, giving an indication of error in estimation of less than $\pm 5\%$. In \figref{Locations}, only the countries listed under New Zealand got a confidence interval ranging between 5% and 8.9%, indicating more expected error in the percentage estimation. Nevertheless, it is still acceptable for some analysis.”

Other Issues: - Figure 3: Please report percentages of stance distributions per language in all the figures.  
*Percentages are now added*

- Please Add a figure showing the distribution of tweets per country after the geo-location annotation- Figure 5: I’m not sure if it is possible to report tweets in the paper without anonymising the usernames due to data privacy and redistribution policies
Twitter terms and conditions allows showing the tweets of individual as long as they didn’t delete it. We checked all the tweets in the Figure, and all of them were still there after 1.5 years, which indicated that they are unlikely to be deleted. Thus we feel it is safe to keep un-anonymized since they add good insights to the analysis. A footnote was added to show that this tweets are still there after more than a year of their post time.

- Many references in the paper are missing the publication year and other information. Please fix. Overall, the paper is well written and structured and targets an interesting and hot research topic. However, there are several issues with the proposed approach and the data annotation that need to be addressed as explained in the above review.


We added this reference and talked about it in the background section.


We think the actual citation is:
@incollection{zhang2011combining,
    title={Combining lexicon-based and learning-based methods for twitter sentiment analysis},
    author={Zhang, Lei and Ghosh, Riddhiman and Dekhil, Mohamed and Hsu, Meichun and Liu, Bing},
    booktitle={HP Laboratories Technical Report},
    year={2011}
}

We added the reference and talked about it in the background section.


We are aware of this report, which addresses supporters of ISIS on Twitter. In the paper we were interested in those who attack/support Muslims.
Reviewer B main actionable comments:

Several aspects of the paper need improvement that the authors need to address to be appropriate for publication. There has been very less work on predicting Islamophobic behavior on social media. However, are attitudes on Islam different from attitudes towards any other topic posing unique challenges for prediction? The authors should better explain this to bring out the novelty of the problem.

*We adjusted the introduction to highlight the contribution and the novelty.*

The unique technical contributions are also not clear. The features are not unique to the problem and can be used for predicting a broad range of user behaviors. The authors should provide conceptual rationale to their features or propose new features which are unique to the problem. Some works which can be useful in this regard are [1,2].

*We added references [1,2] + Add some text. Paper 1 is not relevant to the topic. However, we think the reviewer wanted us to look at the way they explained the usefulness of features. We added an explanation for the rational of using different features in Section 5.3. Paper 2 is on topic and I put a reference to it.*

Comparing with the baselines, such as [3] will help in evaluating the effectiveness of the proposed features. However, there is not such results in the manuscript.

*We commented on paper 3 in the background with text beginning with: “In work closely related to this paper, Qiu et al. …”*

The paper has three tables for different countries (Tables 4, 5, and 6). There are several differences in the results in the three tables that should be explained. For example, the F-measure for the USA using all features is much greater than UK and France.

*The training size for U.S. is larger than UK and UK is larger than France. That’s one of the reasons why the performance is better for US. In case of France, the discussions might be more subtle as the event occurred in France and so there are certainly more variety of opinions and related issues that the model found more difficult to generalize to*
compared to U.S. and UK where the discussions are more general and polarized. The explanations added to section 3.4.

There are lot of typographical errors, for example in the title it should be “Islamophobic” and not “Islamophopic”

Sorry for this mistake. Fixed now