Results

Gensim LDA - Topic Modeling

Gensim LDA[10, 11] gives us the top 10 topics and keywords distribution. We use pyLDAvis [12] to visualize the results of LDA. It gives a distribution and dissimilarity distance in the topics identified by LDA. It also provides a TF-IDF distribution of top 30 keywords associated with the topics. This is represented in the fig 7.1.

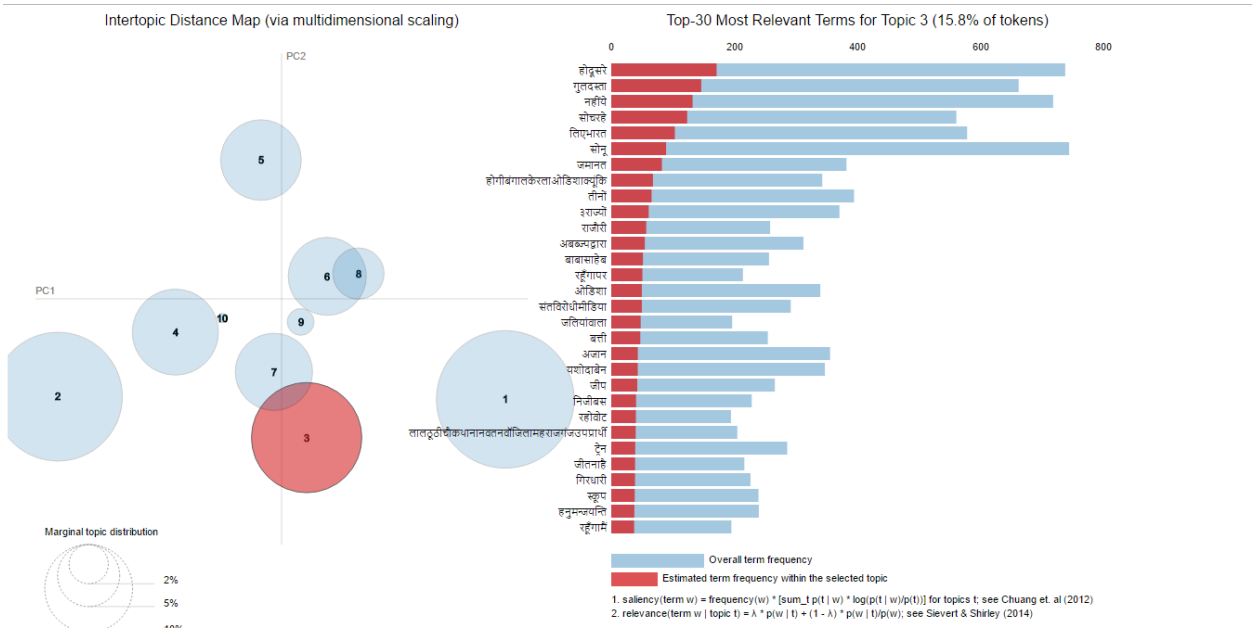

Fig 7.1 pyLDAViz output. Left side gives the topic dis similarity distance, right side gives the tf-idf distribution of top 30 keywords within a selected topic

Below are some of the topics/keywords generated by LDA algorithm

Orignal Topic Keywords	**Word by Word Translation**	**English Translation**
सोनू अजान लाउडस्पीकर फतवे	Sonu Ajaan Loudspeaker Fatwa	SonuNigam talking about Azaan which is call to prayer played on a loudspeaker
होगीबंगालकेरलाओडिशा	willhappenBengalKerala Odissa	Amit shah was on political tour in Indian states of Bengal, Kerela and Odissa
बूचरखाना रोमियोस्क्वाड ऑफिसकी	Slaughter house RomeoSquad Office	Shutdown of slaughter houses and creation of romeo squad in UP
जलियांवाला	Jallianwala baugh	Jallianwala baugh massacre

माल्या	Malya	About Vijay Malya
लालबत्ती	RedLight	Red Beacon
किसानग़रीबीनौकरी	Farmer Poverty Job	Farmers becoming poor and losing their jobs
हनुमानजयंती	Hanuman birth anniversary	Hanuman birth anniversary
ऑस्ट्रेलियाई	Australian	Australian prime minister visits India
वीआईपी	VIP	VIP culture banned in Indian government

Sentiment Analysis

Once we get the topics from LDA, we use these topics to get the sentiments/ opinion of the people tweeting about these topics. For baseline, we use wordnet and our model uses random forest to classify the tweets sentiments. The below graph shows a comparison of the sentiments that we achieved using wordnet and random forest model.

As it is clear from the graphs, random forest out performs wordnet model in almost all the cases. Most of the misclassification of wordnet model is due to the fact that we don't have all the words present in the wordnet, spelling errors, presence of slang words and English words written in hindi.

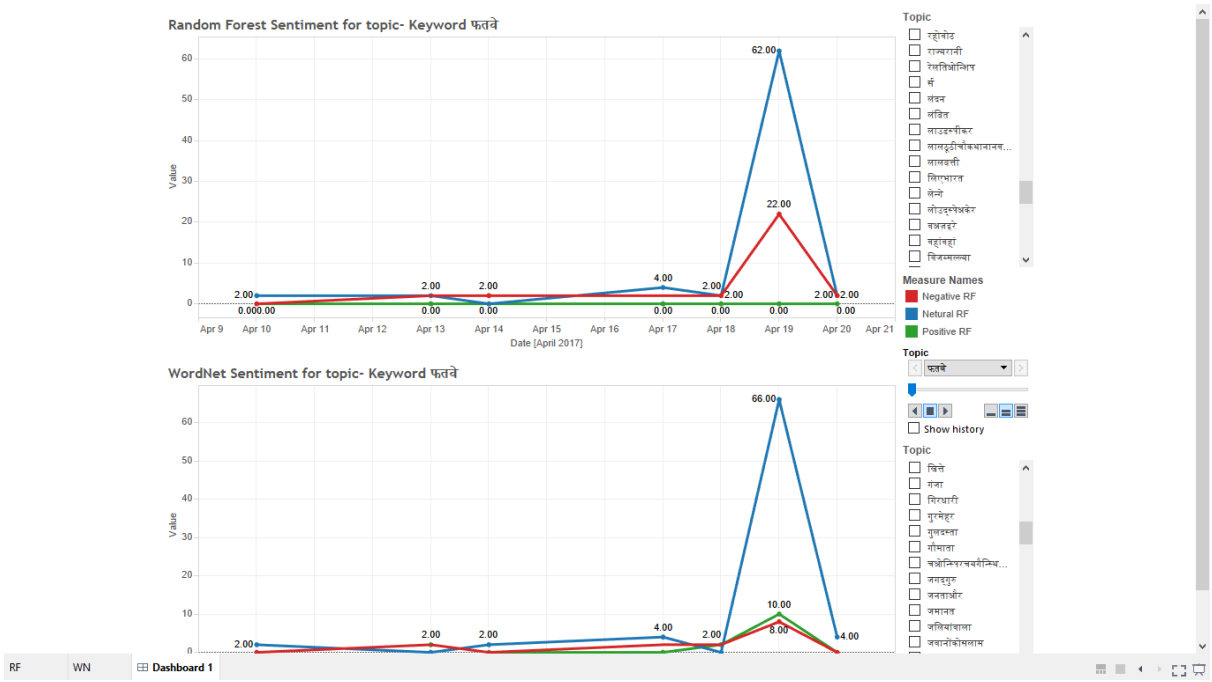

Fig 7.2 Sentiment Comparison for keyword फतवे. (Fatwa)

The first graph represents Random Forest output followed by WordNet output. Red represents negative, green positive and blue as neutral sentiments. We have provided with a checkbox for topic selection and a slider to traverse through the topics. Here, Random forest classifies the data into negative sentiments where as wordnet classifies most of the tweets as neutral. This is due to the fact that we don't have most of the keywords present in the wordnet. This is true for the rest of the figures as well.

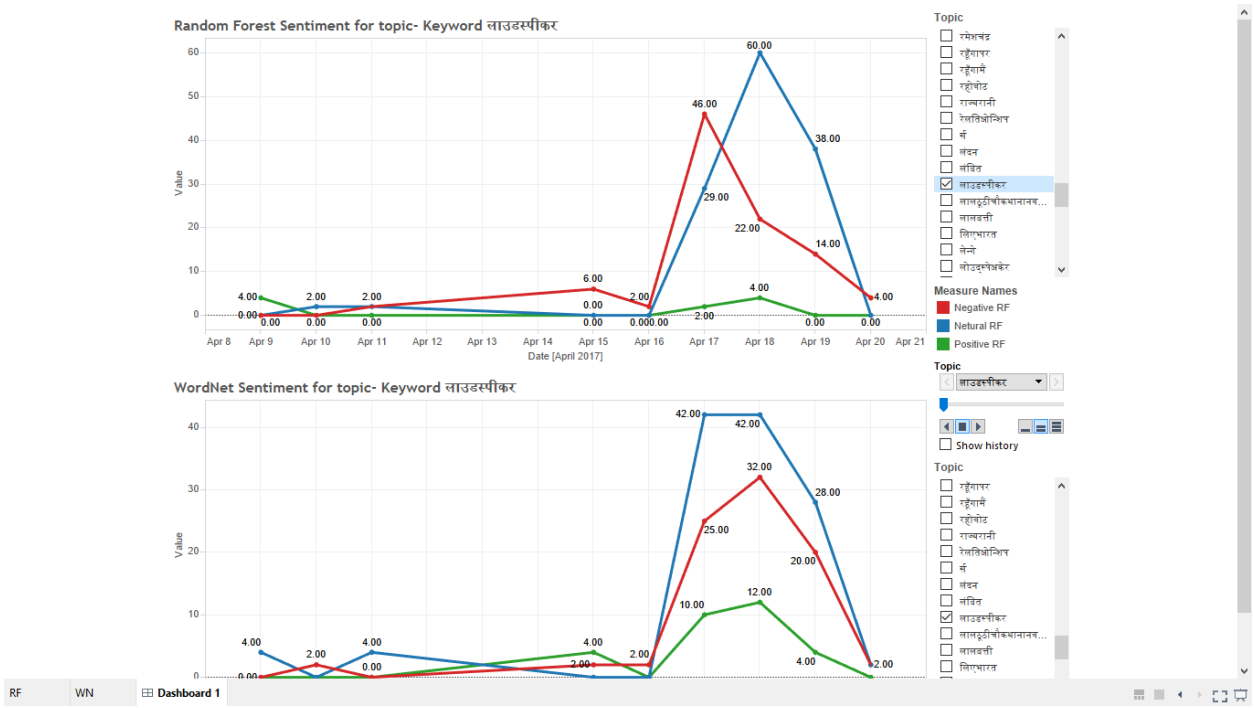

Fig 7.3 Sentiment Comparison for keyword लोउद्स्पेअकेर (Loudspeaker)

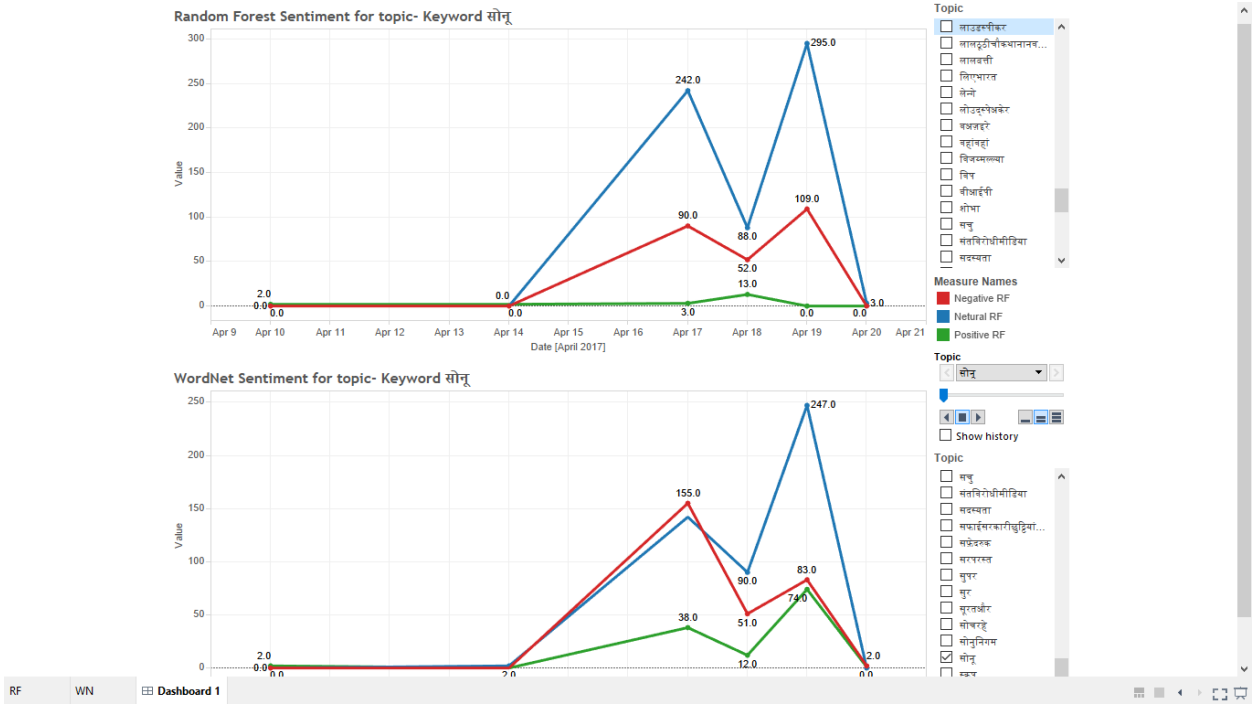

Fig 7.4 Sentiment Comparison for keyword सोनू (Sonu Nigam : Indian Singer).

Transliteration converted sonu nigam in english to (सोनू गम) (Sonu Gum) . WordNet classifies Gum to a negative sentiment as Gum in hindi means “sorrow” hence the overall tweets negative sentiment increases, classifying the tweet to negative sentiment. hence wordnets negative sentiment increases, however people are not actually speaking negative about him. Which is classified correctly by Random Forest.

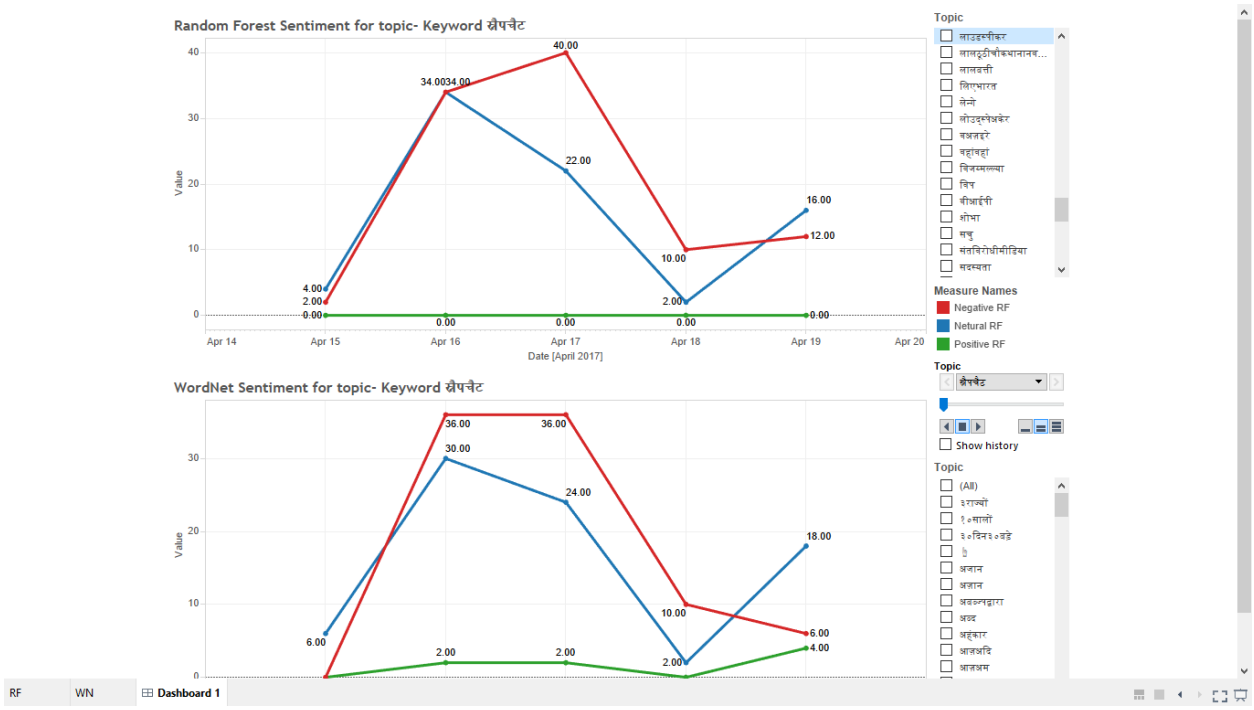

Fig 7.5 Sentiment Comparison for keyword स्नैपचैट(Snachat). Here we don’t see much of a difference in both models