Impact of Sentence Embedding on Sentiment Analysis

Mridul Mishra,
Dept. of Computer Science & Engineering
Parul University, Vadodara, India

Jaydeep Viradiya
Dept. of Computer Science & Engineering
Parul University, Vadodara, India

Abstract: Sentiment Analysis using data-mining or Machine learning is widely used now a days by people and organizations to effectively understand the opinion of people on issues like elections, certain incidents, brands, products, their reaction to situations to gauge the outcome of events like elections, referendums, brand-value, product outcome, success of marketing strategies etc. However, human languages are very abstract in nature and a sentiment analysis model may fail to capture that, and understand that. In this study, we aim to study the results of contextually similar words on a simplistic sentiment analysis model on twitter data, and if there is some flaw in results how can sentence embeddings overcome that.

Index Terms – *Machine Learning, Sentence Embeddings, Sentiment Analysis;*

I. INTRODUCTION

Over the time, social media and their use by people has drastically changed, with more and more people expressing their views, discuss and debate with peers on a large array of topics [1]. This makes these social sites a large pool of real time data.. These posts and discussions are on a variety of data like current affairs, consumer goods, automobiles, politics etc.. As a result, organizations and analysts have started to use the data on social Medias to skim the data according to their use-case like popularity of a leader, to gauge the opinion of people regarding a particular topic. Many brands & people try to position themselves/their products according to these opinions [2]. A sentiment analysis study is done to get a summary of overall sentiment.

These models extract data about a certain topic/product/person and the study the data to get an idea of opinion about that. Here, we will conduct our study, using data on twitter. We created a simple model to extract tweets from twitter about a topic and see the broad positive impact about that topic. And we skimmed the tweets regarding current Prime Minister of India, Mr Narendra Modi using the term "Modi" .

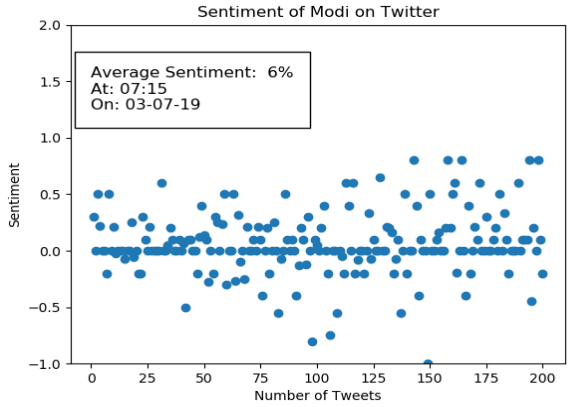

Figure 1Sentiment Result for term "Modi"

The figure above sampled 200 tweets about the keyword "modi" and found average 6% positive sentiment for the topic(the model isn't tuned to high accuracy as the main aim is not sentiment analysis.)
Now ideally, a sentiment analysis model should be able to do :

- **Feature Extraction:** Selecting proper features from the dataset is paramount for the accuracy of the model [3].
- **Capturing similarity:** Based on the phrases/sentences use we try to determine the similarities, e.g search for "mobile battery rating" and "mobile battery power" same information is intended, so we can co-relate this for better results. But on word-level it might be difficult to obtaion co-relation between "power" and "rating" [4].

There are many other important aspects too, but feature extraction is a fundamental requirement for a successful sentiment analysis model. In this study, we aim to explore and understand the significance of the next mentioned point, i.e. capturing contextual similarity.

As mentioned above, a sentiment analysis model should be able to grasp the contextual similarity between different words conveying the same things. A typical sentiment analysis model will fail here, as it isn't trained/designed to capture the similarity between words [5].
To explore this point, we will find the user sentiment on same topic as above, using to different terms, that point at the same topic, the keywords used shall be "pm" (position of Narendra Modi) and "namo"(a moniker used for him on social media). Following results were obtained:

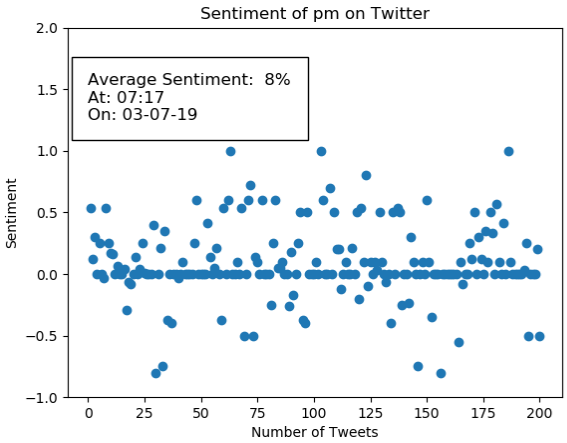

Figure 2Sentiment Result for term "pm"

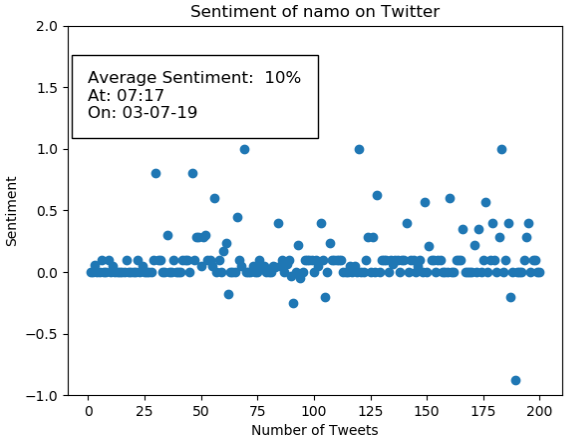

Figure 3Sentiment Result for term "namo"

As is evident from above figures, we can clearly see the variation in results of opinion for different terms that converge to the same topic,

A similar study for Indian Cricket player Mr M.S. Dhoni using different search terms for him, yielded similar deficiency in the model as illustrated by the below figures:

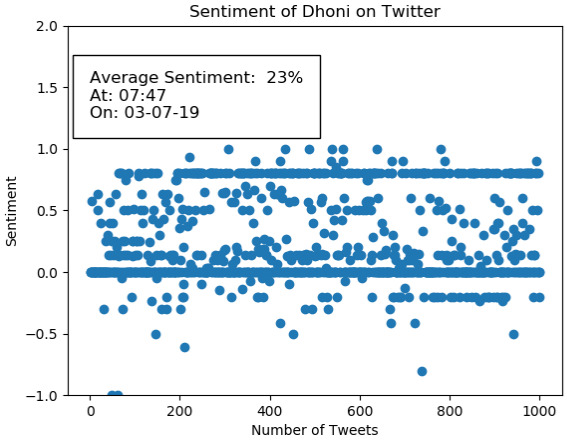

Figure 4Sentiment Result for term "dhoni"

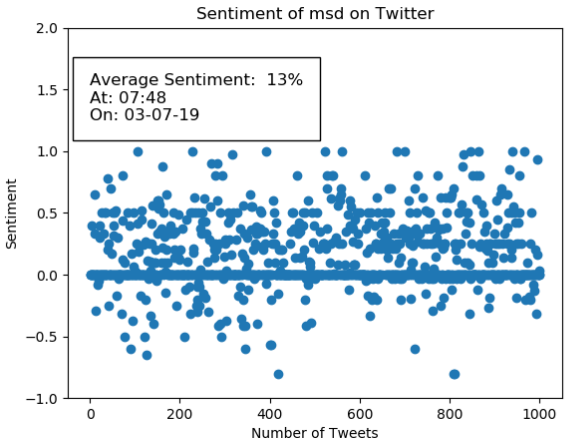

Figure 5Sentiment Result for term "msd"

As is evident from above figures, there is considerable variation in the sentiments regarding the same items when the search terms are changed but are related to the subject. The above illustrations deals with very small dataset to give a proof-of-concept. To practically implement it, we shall be using a large dataset of tweets .

II. CHALLENGE OF CAPTURING CONTEXTUAL SIMILARITY IN SENTIMENT ANALYSIS AND ROLE OF SENTENCE EMBEDDING IN IT.

The results illustrated above clearly show the deficiency of a sentiment analysis model in getting proper sentiment summary for same topics, which may be expressed by different people in different terms [5].

This raises serious questions about the accuracy of sentiment analysis models, which jeopardises the reliability of such study, especially in a region like India, USA which have very high diversity of people thus diversity in expression of words [6].

This is where sentence embeddings come into picture. Sentence embeddings are aimed at capturing semantic, contextual etc similarity between texts [4] [7].

The effectiveness of Sentence Embedding depends on the proximity of their semantic relation and their cosine similarities. It has been observed that the effectiveness prediction of embeddings is affected by i) length of original sentence, ii) order of words in original sentence and iii) words occurring in original sentence [8]. So a model needs to be developed such that it generates sentence embedding for texts to find the similarity/ dis-similarity between texts and based on this the sentiment analysis is done.

In other words the data should first be fed to a sentence embedding model, which shall calculate the embeddings to find the similarities between different pieces of text which are tweets in this case. Then that data should be fed to the sentiment classifier to get the final score.

Theoretically we should see and improvement in the sentiment analysis and overcome the anamoly shown in figure above.

We are in a preliminary stage of our study and believe that successful implementation of these proposed changes should improve sentiment summarization by 5-10%. We will continue our study on this, implement the revised model to find out the results.

III. CONCLUSION

Sentiment Analysis is a widely used tool to understand the pulse of general public on varied topics. Here we have done a very elementary study to understand a basic flaw in sentiment analysis models. In future, we aim to quantify our claims about improvement in performance of sentiment analysis models, by creating a hybrid model of Sentence embedders and sentiment analysis to get improved accuracy.

IV. REFERENCES

1. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2, no. 1–2 (2008): 1-135.
2. Wang, Changbo, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. "SentiView: Sentiment analysis and visualization for internet popular topics." IEEE transactions on human-machine systems 43, no. 6 (2013): 620-630
3. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2, no. 6 (2012): 282-292.
4. Mishra, Mridul K., and Jaydeep Viradiya. "Survey of Sentence Embedding Methods."
5. Tang, Duyu, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. "Sentiment embeddings with applications to sentiment analysis." IEEE Transactions on Knowledge and Data Engineering 28, no. 2 (2016): 496-509.
6. Andugula, Prakash, Surya S. Durbha, Roger L. King, and Nicolas H. Younan. "Domain adaptation approach for classification of high resolution post-disaster data." In 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS, pp. 1733-1736. IEEE, 2013.
7. 4. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." In International Conference on Machine Learning, pp. 1188-1196. 2014.
8. 5. Chen, Yahui. "Convolutional neural network for sentence classification." Master's thesis, University of Waterloo, 2015.