Classification of Tweets Into Facts and Opinions Using Recurrent Neural Networks

Murugan Pattusamy, University of Hyderabad, India*
Lakshmi Kanth, University of Hyderabad, India

ABSTRACT

In the last few years, the growth rate of the number of people who are active on Twitter has been consistently spiking. In India, even the government agencies have started using Twitter accounts as they feel that they can get connected to a greater number of people in a short span of time. Apart from the social media platforms, there are an enormous number of blogging applications that have popped up providing another platform for the people to share their views. With all this, the authenticity of the content that is being generated is going for a toss. On that note, the authors have the task in hand of differentiating the genuineness of the content. In this process, they have worked upon various techniques that would maximize the authenticity of the content and propose a long short-term memory (LSTM) model that will make a distinction between the tweets posted on the Twitter platform. The model in combination with the manually engineered features and the bag of words model is able to classify the tweets efficiently.

KEYWORDS

Classification of Tweets, Long Short-Term Memory Model, Recurrent Neural Networks

INTRODUCTION

Given an option, every individual wants their opinions to be heard and accepted. To accommodate this need, social networking platforms such as Facebook, Twitter, and Telegram, etc. mark their space in the online market. Every platform offers individuals the opportunity to post as much content as they wish. In order to make the post unique, there are high chances that the information shared by the individual will be biased with their opinions than the underlying facts. The need to classify the facts from opinions is therefore essential. The opinions and facts when channelized have got the potential to generate their sentiments. Hence, it is the responsibility of the platform provider to differentiate between facts and opinions to ensure that panic does not prevail in the community (Chatterjee, Deng, Liu, Shan, & Jiao, 2018).

In the past years, the number of people who are active on Twitter has been consistently spiking. Despite having many competitors, Twitter is a widely used marketing tool. In India, even the government agencies have started using the Twitter account as they can get connected to a greater number of people in a short period. Credit to the technological advancements, whatever happens at

DOI: 10.4018/IJTHI.319358

*Corresponding Author

any place on the globe, it gets cascaded to every other part of the globe. With this, there is a plethora of content that is being generated. On an average, every second, around 6000 tweets are emerging, which corresponds to over 3,50,000 tweets per minute, 500 million tweets per day and around 200 billion tweets per year (Hasan, Orgun, & Schwitter, 2019). Interesting insights can be obtained through this data. At the same time, it is desirable to eliminate data points that have opinions. It is crucial that before gaining insights from the tweets, it is beneficial to differentiate the tweets based on their authenticity by considering the person who is tweeting (Deng, Sinha, & Zhao, 2017; Wiebe & Riloff, 2005; Wright, 2009). Dealing with such a humongous volume of data needs much effort. With the advancements of big data technologies and also with the enhanced computational power, dealing with such a variety of data, growing at a rapid pace is convenient. If there is less authenticity in a particular tweet, it may comprise of personal belief or the sentiment of the person.

Understanding both the opinions of the individuals and the facts around the subject has got its business opportunities. In order to tap this potential, the initial step would be to differentiate between the opinions and the facts. The semantics of the tweets should be analyzed before understanding the sentiment of the tweets. After obtaining the sentiment of the tweets, categorize them into their respective classe (opinion or fact). In this research work, the tweets that were related to the airstrike carried out by India in retaliation to the attack on the Indian CRPF soldiers at Pulwama have been considered. This data is analyzed because the situation was panic-driven as the whole of television broadcasting was emphasizing upon this subject.

Moreover, there was an election fever that was picking up in India around the same time. Also, a solution of this sort can be applied to various other instances dealing with varied subject areas. Interestingly, the approach can be extended to other platforms (such as WhatsApp, Instagram) as well.

To address this particular problem statement, the study demonstrates a new algorithm that classifies the authentic tweets from the opinions shared through tweets. In this process, a set of features are manually generated, which enables differentiating the tweets effectively and efficiently. This serves the purpose of supervising the activity that we are performing. These manually engineered features will then be combined with the Bag of Words (BOW) model generated as part of the Natural Language Processing (NLP). After combining the features explicitly, we then use the Long Short Term Memory (LSTM) network, which is an extension of the RNN model (Goodfellow, Bengio, & Courville, 2016). We benchmark the performance of the LSTM network using a labelled dataset (test dataset) and compare its results with other popular and relevant models (Evermann, Rehse, & Fettke, 2017; Ghiassi, Zimbra, & Lee, 2017; Tumasjan, Sprenger, Sandner, & Welpe, 2010; Wang, Wang, Li, Abrahams, & Fan, 2014; Wiebe & Riloff, 2005).

This study makes contributions to (1) understand the importance of distinguishing the authentic tweets from mere opinions shared by the people on the twitter platform, (2) develop a deep learning model by combining the two different types of feature sets to classify the tweets from Twitter, (3) project the best fit model for the given dataset for the purpose of sentiment analysis of the data, and (4) contrast the significance of the finalized model in the present-day situation. It is known that LSTM is good at handling quasi data. Hence, the hypothesis as LSTM is the best model for sentiment analysis is formulated. Though research has been done earlier with LSTM in accordance with textual data. The uniqueness of this work comes from integrating the LSTM model with the BOW features and manually engineered features.

RELATED WORKS

There is a wide range of work that is presently being carried out in this particular domain. Notably, in the last few years in this decade, an enormous number of applications have popped up in this area. These applications can be categorized into the following: event detection, semantic analysis and sentiment analysis. On the whole, much work is being carried out extensively to understand the social media data (tweets in this context) and get the facts right. Interestingly, the problem statements picked up