

Sentiment Analysis of Twitter Data Using Deep Learning



Pranati Rakshit, Pronit Sarkar, Debosmita Ghosh, Shubhankar Roy, Subhadip Talukder, and Partha Sarathi Chakraborty

Abstract In today's world, the use of social networking websites is in the next level. People express their thoughts and opinions on any brand, product or any social events via these sites. Sentimental analysis often uses natural language processing (NLP) to obtain a sentiment behind the text, tweet or comments. In this modern world, sentiment analysis has become one of the most efficient way to mine the public emotions, opinions based on their particular topic of their interest. We have used datasets which consists of tweets from Twitter which contains tweets from various domains. This paper described an approach where a stream of tweets is pre-processed then classified based on the emotion within the text. We used sequential model, with long short-term memory (LSTM) as a layer to train our model. This trained model can be used to analyse any tweets or blogs to obtain sentiments behind that. We have achieved an accuracy of 81.5% from our developed model.

Keywords Sentimental analysis · Tweet data · NLP · Long short-term memory · Accuracy

1 Introduction

Nowadays, social media become an effective path to people for collecting various kind of information [1] as well as a platform for all those people who wants to share their opinions through these sites without any hesitation [2]. One of them is Twitter which has millions of users who always stay active. It is said that at least 500,000 new tweets are posted in Twitter on every single day [3]. Now, Twitter become most used social site for public to share their views, thoughts and opinion on particular topic they are interested in and also Twitter has become a platform for marketing of brands [4].

These social media platforms especially Twitter work as a place to analyse the opportunity for a new product or a business, also as a place to promote or analyse

P. Rakshit (✉) · P. Sarkar · D. Ghosh · S. Roy · S. Talukder · P. S. Chakraborty
Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, India
e-mail: pranati.rakshit@jiscollege.ac.in

existing products to enhance their business [5, 6]. As more and more people are getting connected with the Internet, any product marketed there can be reached to the great number of people. These platforms also generate tons of data from huge amount of users all around the globe on various domains. Data collected by these platforms is also said to be better sample compared to data collected by others means, as they have data on every domain, from all geographical locations, of all age groups. Sentiment-rich data in form of tweets, blogs, comments, reviews, posts, etc. can be used to analyse for the betterment of products and services, which as a result is used for the growth of the business [7].

Sentimental analysis actually analysed the emotions or sentiment behind any blogs, posts, reviews or comments. It usually helps users by gathering all information for a particular product and tell us whether we should buy it or not. This analysis is a huge process of using of pre-processing [8] and text analytics to analysis the sentiment of tweets [9]. Natural language processing plays a major role behind this sentimental analysis [10].

Feature level sentimental analysis is a part of analysis which provides a fine-grained sentiment analysis on certain opinion targets and has wider range of applications on E-business.

Sentimental analysis is also known as opinion mining [11]. It is a very effective tool that is used in any business field or any social media. In any online website, when a company wants to sell a product, it is very important to know about how a customer react about the product for further progress. Sentimental analysis is one among the foremost common ways which analyses an approaching message and tells whether the elemental estimation is positive, negative or unbiased [12]. Sentiment interpretation is one among the toughest challenges within the production of tongue, and sometimes individuals fail to properly analyse emotions (Fig. 1).

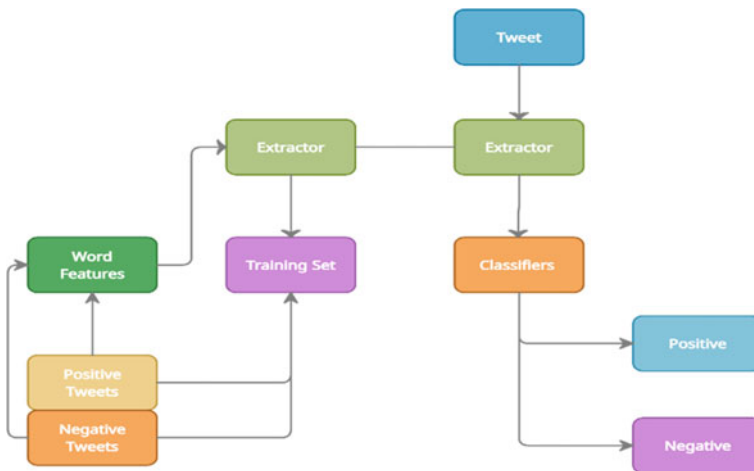


Fig. 1 Sentimental analysis architecture

- (a) Sentimental analysis has some real-life example like reputation management, brand sentimental analysis, etc. Essentially, the application of sentimental analysis gives analysis flexibility and insight into the awarding the organisation and its products.
- (b) Sentimental analysis helps to solve the problem of dealing with huge volume of data. Users can easily find out the emotional tone of any reviews or comments.
- (c) Sentimental analysis helps to mine any data and extract the emotions that underlie social media conversations.
- (d) Sentimental analysis helps to differentiate the positive or negative comments by analysing all data.

Section 2 describes the literature survey, Sect. 3 describes methodology, and result and discussions are depicted in Sect. 4, whereas conclusion is expressed in Sect. 5 with future direction.

2 Literature Survey

From the past few years, we recommended opinion of users, domain experts for making a decision in today's life. According to business changing nowadays, we have to be more effective to our customer's need, e.g. which brand is more popular and effective for the certain product, whether the market value of the product is high or efficient, whether the current series are good or not. According to Appel et al. [13] and Stephen et al. [14], opinion mining can be called as sentimental analysis which plays an important role in this process. It is the objective of emotions like sentiments, expressions that are basic properties in natural language processing by Liu [15], Pak and Paroubek [16], Vinodhini and Chandrasekaran [17], Maks and Vossen [18]. There are some techniques which applied to omit emotions from unstructured data. The analysis of sentiment deals with the identification and classification of opinions or feelings that are present within the source text. According to Zvarevashe and Olugbara [19] and Saberi and Saad et al. [20] within the sort of tweets, status updates, reviews and blog posts, etc., social media generates an enormous amount of sentiment-rich data. In understanding the opinion of the gang, sentimental analysis of this user-generated data is extremely helpful by Feldman [21] and Madhoushi et al. [22]. A path that makes us understand the essential sentiment of the users that they want to express by commenting on our blog by Neethu and Rajasree [23]. Basically, we will be making computer capable of understanding the contents of the text, including the sentiment within them. It will help us to extract all details and observations found in the text and to categorise and arrange the record themselves according to Chowdhury [24] and Indurkhya and Damerou [25]. As a Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches, Twitter sentiment analysis is difficult compared to basic sentiment analysis, thanks to the presence of slang words and wrong spellings by Patel et al. [26], the utmost character limit allowed on Twitter

is 140. For the analysis of feelings from the text, a machine learning approach is often used according to Kharde and Sonawane et al. [27].

3 Methodology

Here, we use sequential model with LSTM as a layer in order to instruct computers to try to learn whatever comes obviously and gain knowledge from encounter. Calculations using machine learning make use of computational strategies to simply “learn” data from data without counting on a preordained condition.

3.1 *Data/Text Cleaning*

In our dataset, there are several types of tweets like comments, review on merchandise, posts concerning one thing, live tweets on matches, etc.; thus, it is important to clean our dataset and each text so as that its square measure usually processed simply. So, cleaning each text was our first initiative. It is done in order to make it easier to process. Filtering out the unwanted symbols, tags, that we do not need so that they do not need to be looked at or processed. At first, we tokenized each sentence in order that it is easier to process. Tokenization is basically the method of splitting or tokenizing a string, text into a list of tokens. Then we clean the text of varied symbols and website marks like @, https, .com, etc. Next, we worked on “stopwords” from each text. In NLP, useless or words with very little meaning are mentioned as stopwords, some commonly used stopwords are “the”, “a”, “an”, “in”. These kinds of words have been programmed to ignore, both when indexing entries for searching and when retrieving them as the results of a query.

3.2 *Split and Train*

While operating with datasets, associate formula of machine learning typically works in two steps. We have a tendency to typically split information around 80–20% train-test ratio. Here, we have around 1.6 million tweets from which we have taken a sample of 5 lac tweets. We train the model using data which are typically referred as training set or training data. Training data is that data which already consists of actual value that must have the ground truth values and thus the algorithm creates changes to the value of the parameters to interpret for training the model. From the 5 lac tweets from the sample, 4 lac tweets are used for training the model and 1 lac for testing purpose.

We can apply the trained model to evaluate any text now on the basis of their score after evaluation by the model and check with the sentimental threshold and classify

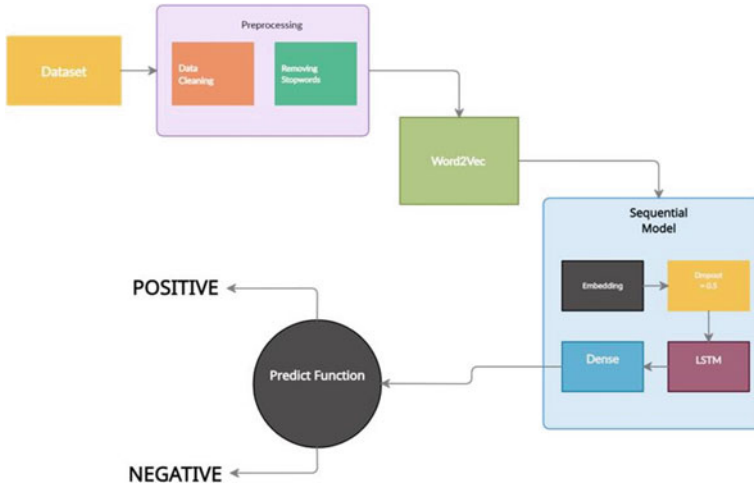


Fig. 2 Workflow model

them as positive, negative or neutral. This model can also be connected in back-end of a blogging site to analyse the sentiments of the blogs automatically, without any interaction (Fig. 2).

3.3 Embedding Layer

As an input to the embedding layer, tweets are given. Any tweet is split into tokens. Each token is transformed into a vector of fixed-sized words, often known as embedding by phrase.

Pre-trained word embedding Word2vec is used for generating the word vector for each token. Among the pre-trained word embeddings, different word embeddings are developed using tweets using the Word2vec algorithm.

The embedding layer output is fixed-sized word vectors that are supplied to the Sequential model as an input for training the model.

4 Result and Discussion

It is very common to remove the stopwords (such as “did”, “doing”, “an”, “the”, “and”) in the pre-processing step as these words tend to have no or very little meaning.

Nevertheless, sometime these stopwords may also have a high impact on the result of certain text. So, to test that, we have conducted an experiment to measure the impact of stopwords removal on the sentimental analysis. We have compared the results of

sequential model with LSTM, both with and without removing the stopwords. It was observed that the model trained with stopwords has outperformed the model trained without stopwords. Hence, we can conclude that removing the stopwords should be avoided, especially for sentimental analysis. We have calculated the accuracy of our model with and without stopwords. After applying the model on our tweet dataset, we have a tendency to get accuracy of seventy nine percent to eighty one point five percent (79% to 81.5%). During this method of analysing the sentiment of tweets, we have used LSTM along of the layers in our ordered layer together with the embedding layer, dropout layer to prevent overfitting, to teach our model. One in every of the foremost uses of LSTM is that this sentimental analysis. The foremost distinction between feedforward neural networks and LSTM is that it is feedback affiliation and it will method entire sequence of data that makes it extremely economical. Sentimental analysis square measure usually machine controlled and should supply resolution to sizable quantity of issues and judgments square measure usually taken on the thought of a giant quantity of information rather than plain intuition that is not correct. It can also be associated integral part of the approach to plug analysis and client service. You may not solely see what individuals believe your own merchandise or services; you may additionally see what your rivals take into account them. With sentiment analysis, the final shopper expertise of your customer's square measure usually discovered simply.

Accuracy (without stopwords)	Accuracy (with stopwords)
79%	81.5%

5 Conclusion

In this present work, we have taken a dataset consisting of tweet data which have various sorts of tweets associated with quite different topics and created a model to analyse each tweet's sentiment. Using this model, we will analyse the other text and determine the sentiment behind it. So, we have got the highest accuracy of 81.5%. As a future direction of this work, we will work with larger dataset and will try to optimise the hyperparameters.

References

1. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of Twitter data. Department of Computer Science, Columbia University, New York, NY 10027 USA
2. Kharde VA. Sentiment analysis of twitter data: a survey of techniques. Department of Computer Engineering, Pune Institute of Computer Technology, Pune University of Pune (India)

3. Liao S, Wang J, Yua R, Sato K, Cheng Z. CNN for situations understanding based on sentiment analysis of twitter data
4. Shah S, Kumar K, Saravanaguru RK. Sentimental analysis of twitter data using classifier algorithms software developer. School of Computing Science and Engineering, VIT University, Vellore, India
5. Adwan OY, Al-Tawil M, Huneiti AM, Al-Dibsi RH, Shahin RA, Abeer A. Twitter sentiment analysis approaches: a survey. Abu Zayed University of Jordan, Amman, Jordan. <https://doi.org/10.3991/ijxx.vx.ix.xxxx>
6. Alayba AM, Palade V, England M, Iqbal R (2017) Arabic language sentiment analysis on health services, pp 114–118
7. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth International conference on computing, communications and networking technologies (ICCCNT), pp 1–5. <https://doi.org/10.1109/ICCCNT.2013.6726818>
8. Haddi E, Liu X, Yong S. The role of text pre-processing in sentiment analysis
9. Kabir AI, Karim R, Newaz S, Hossain MI. The power of social media analytics: text analytics based on sentiment analysis and word clouds on R
10. Shrestha H, Dhasarathan C, Munisamy S, Jayavel A. Natural language processing based sentimental analysis of Hindi (SAH) script an optimization approach
11. Kumar P, Jaiswal UC. A comparative study on sentiment analysis and opinion mining
12. Mehra R, Bedi MK, Singh G, Arora R, Bala T, Saxena S (2017) Sentimental analysis using fuzzy and Naive Bayes. In: 2017 International conference on computing methodologies and communication (ICCMC), pp 945–950. <https://doi.org/10.1109/ICCMC.2017.8282607>
13. Appel G, Grewal L, Hadi R, Stephen AT (2020) The future of social media in marketing. *J Acad Mark Sci* 48(1):79–95
14. Stephen A, Hadi R, Grewal L, Appel G (2019) The future of social media in marketing. *J Acad Market Sci* 48
15. Liu B (2012) Sentiment analysis and opinion mining. *Syn Lect Human Lang Technol* 5(1):1–167
16. Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. *LREc* 10:1320–1326
17. Vinodhini G, Chandrasekaran RM (2012) Sentiment analysis and opinion mining: a survey. *Int J* 2(6):282–292
18. Maks I, Vossen P (2012) A lexicon model for deep sentiment analysis and opinion mining applications. *Decis Support Syst* 53(4):680–688
19. Zvarevashe K, Olugbara OO (2018) A framework for sentiment analysis with opinion mining of hotel reviews. In: 2018 Conference on information communications technology and society (ICTAS). IEEE, pp 1–4
20. Saberi B, Saad S (2017) Sentiment analysis or opinion mining: a review. *Int J Adv Sci Eng Inf Technol* 7:1660–1667
21. Feldman R (2013) Techniques and applications for sentiment analysis. *Commun ACM* 56(4):82–89
22. Madhoushi Z, Hamdan AR, Zainudin S (2015) Sentiment analysis techniques in recent works. In: 2015 Science and information conference (SAI). IEEE, pp 288–291
23. Neethu MS, Rajasree R (2013) Sentiment analysis in Twitter using machine learning techniques. In: 2013 Fourth International conference on computing, communications and networking technologies (ICCCNT). IEEE, pp 1–5
24. Chowdhury GG (2003) Natural language processing. *Ann Rev Inf Sci Technol* 37(1):51–89
25. Indurkha N, Damerau FJ (eds) (2010) Handbook of natural language processing, vol 2. CRC Press
26. Patel AP, Patel AV, Butani SG, Sawant PB (2017) Literature survey on sentiment analysis of Twitter data using machine learning approaches. *IJRST-Int J Innov Res Sci Technol* 3(10)
27. Kharde V, Sonawane P (2016) Sentiment analysis of twitter data: a survey of techniques. arXiv preprint: arXiv:1601.06971