International Journal of Scientific & Engineering Research Volume 11, Issue 8, August-2020
ISSN 2229-5518

Hybrid approach for twitter sentiment analysis using supervised machine learning algorithms

Miss.Nikita Pagar
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
Savitribai Phule Pune University
Pune, India
nikita.pagar16@gmail.com

Prof.B.S.Satpute
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
Savitribai Phule Pune University
Pune, India
b.s.satpute@gmail.com

***Abstract*—Social Media sites like twitter have billions of people share their opinions day by day as tweets. As tweet is characteristic short and basic way of human emotions. So in this paper we focused on sentiment analysis of Twitter data. Most of Twitter's existing sentiment analysis solutions basically consider only the textual information of Twitter messages and strives to work well in the face of short and ambiguous Twitter messages. Recent studies show that patterns of spreading feelings on Twitter have close relationships with the polarities of Twitter messages. In this paper focus on how to combine the textual information of Twitter messages and sentiment dissemination models to get a better performance of sentiment analysis in Twitter data. To this end, proposed system first analyses the diffusion of feelings by studying a phenomenon called inversion of feelings and find some interesting properties of the reversal of feelings. Therefore we consider the interrelations between the textual information of Twitter messages and the patterns of diffusion of feelings, and propose random forest machine learning to predict the polarities of the feelings expressed in Twitter messages. As far as we know, this work is the first to use sentiment dissemination models to improve Twitter's sentiment analysis. Numerous experiments in the real-world dataset show that, compared to state-of-the-art text-based analysis algorithms.**

***Index Terms*—Text Mining, Machine learning, Sentiment analysis, sentiment diffusion, Twitter.**

I. Introduction

Twitter, a popular micro blogging service around the world, has shaped and transformed the way people get information from the people or organizations that interest them[1][4]. On Twitter, users can post status update messages, called tweets, to tell their followers what they are thinking, what they are doing or what is happening around them. In addition, users can interact with another user by replying or republishing their tweets. Since its creation in 2008, Twitter has become one of the largest online social media platforms in the world. Given the increasing amount of data available from Twitter, the polarity of the feelings of mining users expressed in Twitter messages has become a hot research topic due to its wide applications[6]. For example, in analyzing the polarities of Twitter users on political parties and candidates, different tools have been developed to provide strategies for political elections. Commercial companies also use Twitter sentiment analysis as a quick and effective way to monitor people's feelings about their products and brands[1][7][8].

This analysis is done by looking for opinions or sentiments from several sentences or tweets obtained[1]. Therefore, this stack of text data in Twitter is quite valuable because it stores valuable information. To uncover this information, data mining needs to be done using certain techniques. Mining this data can be done using text mining techniques which can be combined also using the Natural Language Pre-processing approach[9]. Furthermore, important data that has been mined needs to be determined by the type of sentiment. This is done by using analytical sentiments. Twitter is one type of social media that is often used. Users use Twitter to convey their Twitter to the general public. The number of Twitter users has reached 330 million people worldwide and every second produces 18000 data. The chirp delivered can be in the form of news, opinions, arguments, and several other types of sentences. This causes twitter to be rich in text that has certain data. In general, someone wants opinions from other people as input to determine decisions. This opinion can be done by asking directly. By asking directly, it takes time and effort to meet people who are believed to ask. Another way is to get opinions from Twitter[1]. Opinions in the form of tweets provided by Twitter with a large amount. However, this opinion must be distinguished based on the type of positive, negative, and neutral opinions. In addition, these tweets have not been grouped according to the categories you want to find. So, it is still widespread and necessary[10][6].

A. Motivation

- Social media, such as Twitter and Facebook, users post many messages including their opinions and feelings. One of the most successful social media, Twitter, allows users to post tweets.
- Twitter sentiment analysis basically only considers the textual information of Twitter messages, but ignores sentiment diffusion information.

II. REVIEW OF LITERATURE

S. Symeonidis, D. Effrosynidis, and A. Arampatzis[1]: Sentiment analysis in microblogging platforms is an essential tool for research and business applications. The analysis of human sentiment and the understanding of human writings by machine learning processes help us to extract useful conclusions about human behavior. Pre-processing is the first step in text Sentiment Analysis, and the use of appropriate techniques can improve classification effectiveness using Linear SVC, Bernoulli Naïve Bayes, Logistic Regression, and Convolutional Neural Networks algorithms. However this paper worked on lemmatization, removing numbers, and replacing contractions techniques and detection accuracy is low.

J. Zhao and X. Gui[2]: This paper discussed the effects of text pre-processing method on sentiment classification performance in two types of classification tasks, and summed up the classification performances of six pre-processing methods using two feature models and four classifiers on five Twitter datasets. However, author worked on static twitter data that's why training performance is low.

X. Zhang, D.-D. Han, R. Yang, and Z. Zhang[3]: In this paper author study the empirical data that crawled from Twitter to describe the topology and information spreading dynamics of Online Social Networks. Propose a measurement with three measures to state the efforts of users on Twitter to get their information spreading, based on the unique mechanisms for information retransmission on Twitter. It is noticed that small fraction of users with special performance on participation can gain great influence, while most other users play a role as middleware during the information propagation. However, removing the incomplete data will cause the loss of information of user profile and user action.

K. Schouten and F. Frasincar[4]: This paper introduced Overview of the state-of-the-art in aspect level sentiment analysis presented in this survey, it is clear that the field is transcending its early stages. While in some cases, a holistic approach is presented that is able to jointly perform aspect detection and sentiment analysis, in others dedicated algorithms for each of those two tasks are provided. Most approaches that are described in this survey are using machine learning to model language, which is not surprising given the fact that language is a non-random, very complex phenomenon for which a lot of data is available. However, this paper introduce state of the art methods on sentiments analysis.

S. Tsugawa and H. Ohsaki[5]: They investigated the relation between the sentiment of a tweet and its virality in terms of diffusion volume and speed by analyzing 4.1 million tweets on Twitter. They used the number of retweets and N-retweet time as measures of tweet virality. They found that negative tweets spread more widely than positive and neutral tweets, and that negative tweets spread faster than positive and neutral tweets when the diffusion volume was large. However, author worked on relation between the sentiment of each tweet and its virality. The relation feature approach is very difficult to calculate.

S. M. Mohammad and S. Kiritchenko[6]: In this paper, compare the performance of several word and character-based recurrent and convolutional neural networks with the performance on bag-of-words. We also investigate the transferability of the final hidden state representations between different classifications of emotions, and whether it is possible to build a unison model for predicting all of them using a shared representation. However, author worked on bag of words techniques.

B. Plank and D. Hovy[7]: This paper focuses on studying two fundamental NLP tasks, Discourse Parsing and Sentiment Analysis. The development of three independent recursive neural nets: two for the key sub-tasks of discourse parsing, namely structure prediction and relation prediction; the third net for sentiment prediction. However, this work is carried out manually so it is time consuming and expensive.

S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin[8]: In this paper explored an application of deep recurrent neural networks to the task of sentence-level opinion expression extraction. DSEs (direct subjective expressions) consist of explicit mentions of private states or speech events expressing private states; and ESEs (expressive subjective expressions) consist of expressions that indicate sentiment, emotion, etc., without explicitly conveying them. However, Mainly Time consuming and resource consuming for the system.

J. Bollen, H. Mao, and X.-J. Zeng[9]: In this paper analyse electoral tweets for more subtly expressed information such as sentiment (positive or negative), the emotion (joy, sadness, anger, etc.), the purpose or intent behind the tweet (to point out a mistake, to support, to ridicule, etc.), and the style of the tweet (simple statement, sarcasm, hyperbole, etc.).There are two sections: on annotating text for sentiment, emotion, style, and categories such as purpose, and on automatic classifiers for detecting these categories.

S. M. Mohammad and P. D. Turney[10]: In this paper, investigate whether public mood as measured from large-scale collection of tweets posted on twitter.com is correlated or even predictive of DJIA values. The results show that changes in the public mood state can indeed be tracked from the content of large-scale Twitter feeds by means of rather simple text processing techniques and that such changes respond to a variety of socio-cultural drivers in a highly differentiated manner. However, first drawback is Low barrier to creating accounts and second is Weak defences, slow response.

III. PROPOSED METHODOLOGY

Proposed sentiment diffusion on Twitter by investigating sentiment reversal, the phenomenon that a tweet and its retweet have different sentiment polarities. We analyze the properties of sentiment reversals, and propose a sentiment reversal prediction model.

To predict the sentiment polarity of each Twitter message, we propose an iterative algorithm called SentiDiff, which takes the inter-relationships between textual information of Twitter messages and sentiment diffusion patterns into consideration. Given a tweet and its retweet, if their sentiment polarities predicted by textual information based sentiment classifier are consistent with the prediction result of sentiment reversal, the probability of messages to be classified correctly by textual information based sentiment classifier will increase. Otherwise, the probability will decrease. In this way, sentiment reversals can be combined with textual information of Twitter messages.

A. Architecture

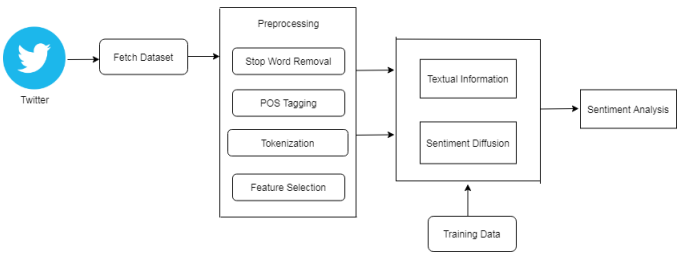

Fig. 1. Proposed System Architecture

B. Algorithm

1. SentiDiff Algorithm

In this section, we propose an iterative algorithm, called SentiDiff, to combine textual and sentiment diffusion information in a supervised learning algorithm. Before going to the details, we first list the notations used in this section in Table

Notation	Meaning
$child(m_i)$	Child tweets of Twitter message m_i
$parent(m_i)$	Parent tweet of Twitter message m_i
$TL(m_i)$	Sentiment label of Twitter message m_i predicted by textual information based sentiment classifier
$TP(m_i, sl)$	Probability of Twitter message m_i to be classified with sentiment label sl correctly by textual information based sentiment classifier
$SP(m_i, m_j)$	Probability that sentiment reversal occurs between Twitter messages m_i and m_j predicted by sentiment reversal prediction model
$TSP(m_i, sl)$	Probability of Twitter message m_i to be classified with sentiment label sl after combining textual and sentiment diffusion information
$FL(m_i)$	Sentiment label of Twitter message m_i after combining textual and sentiment diffusion information

First, we train textual information based sentiment classifier and sentiment reversal prediction model based on the labeled dataset. Then, given a new set of Twitter messages which reside in the same cascade tree, the sentiment polarity of each Twitter message is predicted as sentidiff Algorithm. The basic idea of fusing textual and sentiment diffusion information in SentiDiff algorithm is that if the prediction results between textual information based sentiment classifier and sentiment reversal prediction model are conflicting, the probability of Twitter messages to be classified correctly by textual information based sentiment classifier will decrease. Otherwise, the probability will increase.

We split our tweet and retweet data with sentiment labels into training, validation and test sets. The training set is used to train textual information based sentiment classifiers and sentiment reversal prediction model, the validation set is used for testing the generalization performance of sentiment classifiers and sentiment reversal prediction model, and the test set is for blind evaluation. The percentage of dataset used as the training set is indicated by a variable pc. For example, pc = 0.7 indicates that 70 of overall repost cascade trees are treated as training set. Then half of the remaining cascade trees are used for validation set, and the rest of the cascade trees are for test set.

2. Random Forest Algorithm

Step 1: Let the number of training cases be N, and the number of variables in the classifier be M.

Step 2: The number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.

Step 3: Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.

Step 4: For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

Step 5: Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).
For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

C. Mathematical Model

The mathematical model for combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis is as-

$$S = \{I, F, O\}$$

Where,
I = Set of inputs
The input consists of set of twitter tweets and retweets.
F = Set of functions

$$F = \{F1, F2, F3\}$$

F1: Textual Information

$$TweetContents$$

F2: Repost Cascade Tree
Repost cascade tree is a directed, acyclic labeled graph, which is used to capture the relationships between a tweet and its retweets.

$$T(V, E, l)$$

Where,
V – set of nodes,
E – Set of edges,
L - Functions
F3: Repost Diffusion Network
Repost diffusion networks to describe how users interact with each other on Twitter

$$N(V, E)$$

Where,
V – set of nodes,
E – Set of edges,
F4: Sentiment Reversal
Sentiment reversal is defined as the phenomenon that a tweet (parent tweet) and its retweet (child tweet) have different sentiment polarities.

$$l(i) \neq l(j)$$

Where,
l – Functions
i – Parent tweet
j - Child tweet

O=Sentiment Analysis (i.e. Positive, Negative, Neutral)

IV. Results and Discussion

The section shows overall accuracy of SentiDiff and Random Forest classification technique . So this works gives better sentiment analysis results compare to existing method. The experimental result validation,
TP: True positive (correctly predicted number of instance)
FP: False positive (incorrectly predicted number of instance),
TN: True negative (correctly predicted the number of instances as not required)
FN false negative (incorrectly predicted the number of instances as not required),
On the basis of this parameter, we can calculate four measurements
Accuracy = TP+TNTP+FP+TN+FN
Precision = TP TP+FP
Recall= TPTP+FN
F1-Measure = 2PrecisionRecall Precision+ Recall.

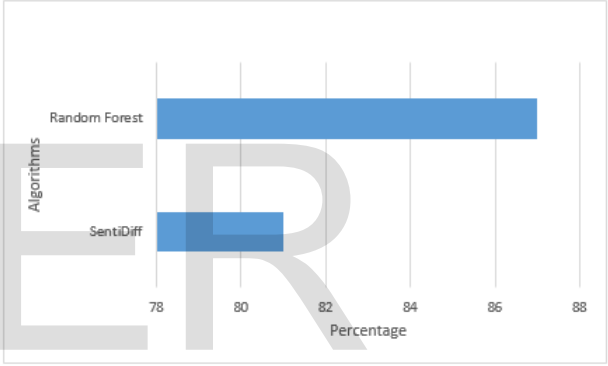

Fig. 2. Accuracy Graph

	Existing System(SentiDiff)	Proposed System(RF)
Accuracy	81.29	87.26

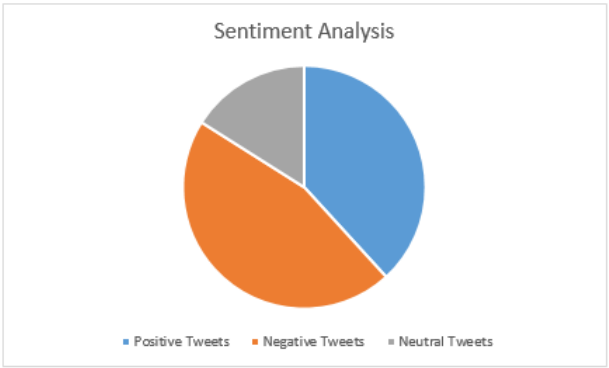

Fig. 3. Accuracy Graph

Sentiments	Classification Results
Positive Tweets	81
Negative Tweets	97
Neutral Tweets	34

V. Conclusion

Polarity of mining sentiments expressed in Twitter messages is a significant and challenging task. Most existing Twitter sentiment analysis solutions consider only the textual information of Twitter messages and cannot achieve satisfactory performance due to the unique characteristics of Twitter messages. Although recent studies have shown that patterns of feeling diffusion are closely related to the polarities of Twitter messages, existing approaches are essentially based only on textual information from Twitter messages, but ignore the dissemination of information about feelings. Inspired by the recent work on the fusion of knowledge of multiple domains, take a first step towards combining textual information and spreading feelings to get a better performance of Twitter's sentiment analysis.

References

[1] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," Expert Systems with Applications, 2018

[2] J. Zhao and X. Gui, "Comparison research on text pre-processing methods on twitter sentiment analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.

[3] X. Zhang, D.-D. Han, R. Yang, and Z. Zhang, "Users participation and social influence during information spreading on twitter," PloS one, vol. 12, no. 9, p. e0183290, 2017.

[4] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 813–830, 2016.

[5] S. Tsugawa and H. Ohsaki, "Negative messages spread rapidly and widely on social media," in Proceedings of the 2015 ACM on Conference on Online Social Networks. ACM, 2015, pp. 151–160.

[6] S. M. Mohammad and S. Kiritchenko, "Using Hashtags to Capture Fine Emotion Categories from Tweets," Computational Intelligence, vol. 31, no. 2, pp. 301–326, 2015.

[7] B. Plank and D. Hovy, "Personality Traits on Twitter —or— How to Get 1,500 Personality Tests in a Week", Proc. of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2015, pp. 92–98.

[8] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets" Information Processing and Management, vol. 51, no. 4, pp. 480–499, 2015.

[9] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," J. of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.

[10] S. M. Mohammad and P. D. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," in Proc. of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. ACL, 2010, pp. 26–34.