Tweets Mining for Classification and Rapid Response for Pessimistic Ones

R. Renuga, K. Reshma

Abstract: ***Social media is virtual network of communities/groups where people can create and share media /opinions/ideas on various things/ objects/ persons/ topics related to entertainment, sports, politics, science, travel etc. Twitter is the most popular micro-blogging site, which provides access to the data for non-commercial and research purpose. Mining topics in Twitter is increasingly attracting more attention. Our aim is to automate the process of responding to pessimistic tweets in feedback given to electronic products. In this way customer will be satisfied by quick response from company and retailer will get some time to work on the customer problem. So Customer-Retailer relationship can be improved. In this paper we present how Open source social media intelligence (OSSMInt) can be applied on twitter tweets to extract necessary information from the feedback that has been tweeted by customers of a particular product and quick response to those customers who are dissatisfied with product. Here we concentrate on negative feedback and satisfy the customer temporarily by allowing the retailer to work on the product. We use necessary algorithms, tools and techniques to get the data from twitter, classify it and reply the customer with quick response who has given negative feedback. This classification can be shared with a data scientist for further procedure to get a solution to the problem escalated by the customer. This works as a connecting bridge between a customer and retailer.***

I. INTRODUCTION

So many companies are now a days are encouraging its customers to give feedback over social media, and they are having a customer consultant team who will constantly check the social media and escalate the feedback and the respective complaints in it to the corresponding team, who will check the issue and solve at earliest.

But this needs a lot of manual work and patience, we thought of automating the first part of it, thereby aiding the betterment of consumer-retailer relationship.

TENTATIVE PROCEDURE

1. Getting data from twitter with particular hashtags (relevant to the product/service/company).
2. Identifying the sentiment of the tweet. (Using a pre-trained model).
3. Send response immediately if sentiment turns out to be negative.

II. LITERATURE REVIEW

The main focus of this project is to get sentiment analysis of text in tweet. We did literature survey on existing sentiment analysis algorithms.

Investigating the Role of Twitter in E-Governance by Extracting Information on Citizen Complaints and Grievances Reports

Nowadays social media, such as: Twitter, Facebook and Instagram, has become a necessity of everyone's life. There has been a noticeable adoption of social media and upward trend in its usage by government agencies for not just disseminating information but also acquiring information such as complaints and grievances from citizens. Evidence shows that twitter is the most widely used micro blogging platform in Internet. Based on the analysis made on several Indian Government Twitter accounts, authors find that an active government twitter handle receives an appropriate of 5 tweets per minute. They conducted a literature survey in the area of complaints and grievances identification on social media and divide the existing literature into two lines of research area:

1. Usage of twitter micro blogging website to report and complaint and grievances.
2. Mining public complains and communication on twitter for building prediction models for situation awareness.

Many researchers have come up with different applications of twitter such as: application that automatically determines road hazards, methodology to mine tweet texts to extract incident information on both highways and traffic related data, a real time monitoring system for traffic event detection from twitter stream analysis, a solution for real time identification of small scale incidents using micro blogs, a study on social based traffic information extraction and classification consisting of mining twitter for traffic congestion and etc.

In this paper authors concentrated on how to develop a solution to automatically resolve the challenges of manual inspection and to build a text analysis-based model to address the NLP challenges in micro posts. They investigated text classification techniques for automatically identifying com- plaints tweets and assigning them to predefined labels based on the topic of content. The problem of automatic identification of citizens' complaints has been formulated as a one-class classification problem. Authors proposed a text classification-based approach consisting of various components performing several tasks: tweets extraction from public agencies account, enrichment and enhancement of raw micro posts, learning the features of non-compliant and complaint report tweets, developing a baseline classification approach, use of ensemble techniques to improve the baseline method, empirical analysis and performance evaluation. They pre- processed the sampled data sets and addressed the challenge of noisy content in the tweets.

Revised Manuscript Received on December 30, 2018.

R. Renuga, Assistant Professor, Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu.

K. Reshma, Assistant Professor, Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu.

The proposed algorithm performs syntactic enhancement of the tweets and consists of five phases primarily named as hash tag expansion, @user- name mention expansion, spell error correction, acronym and slang treatment and sentence segmentation.

They also observed that not all tweets posted on these accounts are complaint reports and rather are either off-topic or discussions not relevant to the complaint department and divided such tweets into 4 categories (AISP): Appreciation posts, Information Sharing and Promotional tweets. Appreciation tweets , posts made by citizens for praising the government for their work or resolving their previous complaints, Information Sharing tweets tweets posted by users to share daily news about the events or policies initiated by the government, Promotional tweets, tweets posted by a different official account of same public agencies. They used the AISP tweet classifier method proposed in their previous study and classify AISP tweets and filter the unknown posts that may or may not be a complaint report about killer roads.

For case study 1, they used an ensemble learning based Support Vector Machine (SVM) classifier as the next step of processing pipeline of solution method and used the feature vector model created in previous phase and learn their one-class classification model-a tweet either belongs to complaint and grievance class or identified as unknown and identified that the performance of an SVM classifier can be improved by modifying the kernels of the classifier. To investigate the performance of proposed approach and evaluate the performance across various dimension, they trained their model by varying the kernel parameter of SVM: linear, polynomial and Radial Basis Function (RBF) Kernels. They use Alchemy Concept API by IBM Watson and identify the hidden topics in complaints. They address the challenges of keyword spotting methods and use NLP based methods to find such words and label these complaints into the most likely topic and subtopic defined in the taxonomy hierarchy. According to empirical results, linear kernel in SVM out- performs RBF kernel with a reasonably high margin with accuracy up to 60%.

For case study 2, based on the inspection of complaints reported posted to official account, they divide tweets into 3 categories: Useful tweets, nearly useful tweets and Irrelevant tweets and used Rule-based classifier trained on the features extracted. Irrelevant tweets , if the post is not about the poor conditions and irregularities of roads or highways causing life risks or poor experience to the citizens, Useful tweets, post which is a clear indicator of complaints and can be used to identify the low-level details of the issue faced by citizens, Nearly-useful tweets, tweet posted for complaining a report but containing incomplete or insufficient information about the issue. Based on the results acquired by rule-based classifier, they classified bad road complaints with an overall accuracy of 67%.

VADER

They have come up with new rule-based model for sentiment analysis of social media text called VADER (Valence Aware dictionary for sentiment reasoning, and they benchmarked their results with other gold standard lexicons in both polarity based like LIWC (Linguistic Inquiry and Word Count), GI (General Inquirer), Hu-Liu04, and valence based like ANEW (Affective Norms for English words), SentiWord Net - SWN (NLTK package of python), SenticNet - SCN. Polarity based - lexicons in which words are categorized into binary classes like positive and negative classes. Valence based - lexicons in which words are associated with valence scores for sentiment intensity.

They have mentioned that all of the above-mentioned gold standard sentiment lexicons doesn't perform well over modern, dynamic and largely available and evolving social media and micro blogging text which may have modern day shortcuts like ROFL, WTF etc. and also, the emoticons (smiley symbols) etc. They also have sentiment attached to it, which will alter the total sentiment of the text. To bridge this gap, they came up with a solution VADER, and they benchmark edit over other sentiment lexicons. Machine Learning Approaches and their disadvantages: -

1. Naive Bayes Classifier
2. Maximum Entropy
3. Support Vector Machine (classification and regression models)

They need training data which are, as validated with sentiment lexicons sometimes are difficult to acquire. They also depend on training data to represent as many features as possible. They are often computationally expensive and memory intensive. Sometimes features derived by them are not human interpretable (Neural Networks). Advantages of approach proposed.

1. It will work well on modern day social media style text that contains shortcuts and emoticons), can also be extended to various domains.
2. It requires no training data, but is constructed from a generalize, valence-based, human curated gold standard sentiment lexicon.
3. It is fast enough to use on streaming data. Many of the social media (like Twitter, Facebook, Reddit etc.) provide tools called API- Application Programming Interface through which researchers and developers can access their data.
4. It doesn't severely suffer from speed-performance trade off.

PROPOSED APPROACH SUMMARY STEPS

1. Examine all lexical features of existing well-established and human validated sentiment lexicons (LIWC, GI, ANEW).
2. Supplement with additional lexical features commonly used to express sentiment in social media and micro-blogging websites (emoticons, acronyms, slang).
3. Use wisdom of crowd approach to establish point estimations of sentimental valence for each 9000+ lexical feature candidates.
4. Kept 7500 lexical features w/mean valence zero, and SD <=2.5 as a human validated gold-standard sentiment lexicon.
5. Use data driven iterative inductive coding analysis to identify generalize heuristics for assessing sentiment in text.

6. Evaluate the impact of grammatical and syntactical rules on perceived sentiment intensity of text.
7. Establish ground truth point estimates of sentiment valence on corpora (data set containing texts) rom four distinct domains using aggregate data from multiple human raters.
8. Compare VADER sentiment analysis to 11 baselines: LIWC, GI, ANEW, Hu-Liu04, WSD, SWN, SCN, NB, ME, SVM-C and SVM-R.

III. METHODOLOGY

MODULE 1: COLLECT DATA FROM TWITTER

In order to have access to Twitter data programmatically, we need to create an app that interacts with the Twitter API. The first step is the registration of your app. In particular, you need to point your browser to http://apps.twitter.com, log-in to Twitter (if you're not already logged in) and register a new application. You can now choose a name and a description for your app(for example(Mining Demo or similar). You will receive a consumer key and a consumer secret: these are application settings that should always be kept private. From the configuration page of your app, you can also require an access token and an access token secret. Similarly to the consumer keys, these strings must also be kept private: they provide the application access to Twitter on behalf of your account. The default permissions are read-only, which is all we need in our case, but if you decide to change your permission to provide writing features in your app, you must negotiate a new access token. Twitter provides REST APIs you can use to interact with their service. There is also a bunch of Python-based clients out there that we can use without re-inventing the wheel. In particular, Tweepy in one of the most interesting and straightforward to use. Tweepy provides the convenient Cursor interface to iterate through different types of objects. The JSON response from the Twitter API is available in the attribute _json (with a leading underscore), which is not the raw JSON string, but a dictionary.

In case we want to "keep the connection open", and gather all the upcoming tweets about a particular event, the streaming API is what we need. We need to extend the *StreamListener()* to customize the way we process the incoming data.

Depending on the search term, we can gather tons of tweets within a few minutes. This is especially true for live events with a world-wide coverage (World Cups, Super Bowls, Academy Awards, you name it), so keep an eye on the JSON file to understand how fast it grows and consider how many tweets you might need for your tests.

MODULE 2: WRITE DATA TO MYSQL USING MYSQLDB

MySQLdb is an thread-compatible interface to the popular MySQL database server that provides the Python database API. Before connecting to a MySQL database, make sure of the followings -

1. You have created a database TESTDB.
2. You have created a required table in TESTDB.
3. This table required fields of your data that you streamed from twitter.
4. Python module MySQLdb is installed properly on your machine.
5. You have gone through MySQL tutorial to understand MySQL Basics.

If a connection is established with the data source, then a Connection Object is returned and saved into db for further use, otherwise db is set to None. Next, db object is used to create a cursor object, which in turn is used to execute SQL queries. Finally, before coming out, it ensures that database connection is closed and resources are released. Once a database connection is established, we are ready to create tables or records into the database tables using execute method of the created cursor.

MODULE 3: PREPOSSESS DATA AND WRITE INTO CLASSIFIER

Once we have collected tweets and stored them in MySQL there will be many key attributes to text like text, Created_at, favorite_count, re-tweeted, lang, id, place, user, entitles. From these attributes we take the required ones and write into classifier using MySQLdb.

MODULE 4: CLASSIFY USING RANDOM FOREST

Random Forest is an ensemble of decision tree classifiers which will output a combined prediction value of each tree in the ensemble. Each decision tree is constructed by using a random subset of the training data with a fixed probability distribution. The deeply grown decision trees have low bias and high variance. Hence they can learn irregular patterns and over fit their training sets. Random forests give improvement over just bagged trees because they decorrelates the trees in the Random forest. The decorrelation is achieved while building RF. While building the decision trees for RF, a random sample of some M training samples are chosen as split candidate from the original N training set, during a split in a tree is happens each time. This strategy is good because, if at all any strong training sample to split and subsequently all trees look similar and they become correlated and any reduction in the variance is only average of many correlated predictions. But RF each split will consider subset of training samples and hence on an average (N-M)/M splits not even consider that strong training sample and so other training samples also have more chance to be splitting candidates. And due to this procedure the generated trees will become decorrelated and average of the results of the classifier trees will be less variance and hence gives more reliable solution.

If RF is built with M=N samples then it is as good as Bagging tree. But generally RF is built with M = sqrt(N) where N is size of training sample and reduces both test error and out-of-bag error.

MODULE 5: REPLY FOR PESSIMISTIC ONES

Based on the output from Random Forest tweets are classified as positive and negative sentiment on the product. If response is negative, send quick response by using ID retrieved from twitter

IV. CONCLUSION

There were few works on twitter handling, getting data from twitter for the analysis purpose. Support Vector Machine (SVM) algorithm was used by them whereas we have used Random Forest algorithm which has shown the better results when compared with the other algorithms.

In our experiments we have successfully done identification of pessimistic tweets regarding electronic products. In future, we try to filter the tweets and predict to which department this tweet must be sent. This will be done purely based on the words that have been used by the tweeters.

REFERENCES

1. Swati Agarwal., Ashish Sureka.: Investigating the Role of Twitter in E-Governance by Extracting Information on Citizen Complaints and Grievances Reports.
2. Agarwal, S., Mittal, N., Sureka, A.: Potholes and bad road conditions-mining Twitter to extract information on killer roads. In: The ACM India Joint International Conference (CoDS- COMAD), India. ACM (2018, under-review)
3. Agarwal, S., Sureka, A.: Using KNN and SVM based one- class classifier for detecting online radicalization on Twitter. In: Natarajan, R., Barua, G., Patra, M.R. (eds.) ICDCIT 2015. LNCS, vol. 8956, pp. 431-442. Springer, Cham (2015).
4. Agarwal, S., Sureka, A.: Investigating the potential of aggregated tweets as surro- gate data for forecasting civil protests. In: Pro-ceedings of the 3rd IKDD Conference on Data Science, p. 8. ACM (2016)
5. Mittal, N., Agarwal, S., Sureka, A.: Got a complaint?- Keep calm and tweet it!.In: Li, J., Li, X., Wang, S., Li, J., Sheng, Q.Z. (eds.) ADMA 2016. LNCS (LNAI), vol. 10086, pp. 619-635. Springer, Cham (2016).
6. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proc. WLSM-11s*. Davidov, D., Tsur, O., Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *ICCL-10.*
7. De Smedt, T., Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*, 2063-2067.
8. Fellbaum, C. (1998).*WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
9. C.J. Hutto.,Eric Gilbert.: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.