Assignment 3: Take Home Exam

31005 Machine Learning Spring 2019

Name: Joshua Gavin Valerie

Student ID: 13022658

Research Question:

Question 1 (Methods and challenges in predicting how the user can be converted to change his/her support, as well as ethical considerations in social media).

Context:

Data Analyst in polling organisation tasked to use messages from social media to analyse how user can be converted to change his/her support.

Github Link

Jupiter Notebook pdf Report Link

"https://github.com/joshuagavinv/UTS_ML2019_ID13022658/blob/master/assigment_3_take_home_exam.ipynb"

Microsoft word pdf Report Link

1. Introduction

Social Media Technologies have given raise the growth of User Generated Content, promoting personal belief, free review/opinion platform such as Twitter, Facebook (Stieglitz et al 2018). As messages composed of massive varieties of unstructured, subjective data/ opinionated information, sentiment analysis can be used to detect the positive/negativity of message expression to determine the attitude of a person towards products, campaign candidate and etc (Gomes & Casais 2018). Social Media Analytics consist of 3 steps which are, data collection, preparation and analysis which was research to be the most used approach in IS(Stieglitz et al 2018). In comparison with traditional opinion polling survey, social media allows people to freely state their feelings in social media, which produce large public data used for research within lower cost and time period (Karami Bennet & He 2018).

2. Challenges

Although, social app contains rich unstructured emotional information that can represent the true sentiment of the society, social media characteristic will potentially bias the analysis results. The following are challenges mentioned from literature discussion:

2.1. Veracity of social media messages

As social media it's a public platform widely used by everyone, the challenge is in determining whether the message is relevant or not. For instance, the high volume data can be contributed by spam messages which increased the amount of data making analysis unreliable (Stieglitz et al 2018). Furthermore, rumours, hoax, can have a detrimental influence to other users which in turn bias the results (Stieglitz et al 2018). As such, determining the credibility of users and message itself will be crucial in deriving useful implication. Based on Stieglitz et al research (2018), the solution was to filter untrustworthy information by investigating user and spam detection. Another challenge is the handling of missing data in messages such as GPS-tag. As the naïve approach will be to exclude them it is researched that only 1-2% of twitter have GPS-tag information, meaning that we'll lose lots of information, if we continue with this approach (Stieglitz et al 2018). Alternatively, predicting the missing value is possible, but the reliability of this is still in question (Stieglitz et al 2018).

2.2. Usage of emoji, figurative, informal word

As social media such as Twitter with 140 character limits, established boundaries for length of posts, it's not informative and contain a lot of noises such as articles and stop words, which will worsen the prediction (Zhang, Xu & Jiang 2018). Furthermore, what make this limitation, difficult to address is the variability unstructured sentences can have (Stieglitz et al 2018). Emoji which can also determine feelings are considered as noise, in traditional opinion mining, due to it's non-text content, however as the popularity rises it can't be ignored anymore (Zhang, Xu & Jiang 2018). Another issues is metaphor, which contain hidden meanings, understandable by human, but might lead to different sentiment prediction by machine as irony/sarcasm can twist the message polarity (Suliz et al 2016). As Social media often used by youth as well,

intentional jargon, typo such as okaii, OMG might not be correctly understand as these area evolve rapidly. Possible solutions for figurative and informal is mapping them to a knowledge graph/storage however as it's difficult to construct and maintain it's currently not applicable due to rapid growth of terms (Zhang, Xu & Jiang 2018).

2.3. Sentiment Scoring for context-dependent words

As each post will be evaluated by their sentiment score, each of the word will be evaluated. However some terms are dependent on the context they are in (Zhang, Xu & Jiang 2018). For instance, heavy laptop and heavy investments will have different polarity, negative and positive respectively. As machine learning approach is not informative in telling whether the issue is solved the authors proposed a topic lexicon/dictionary approach, however the other challenge is users might change into a different topic in one posts, which can't be solve with this approach (Zhang, Xu & Jiang 2018).

3. Proposed System

3.1. Existing Approach

Sentiment analysis prominently divide into lexicon and machine learning. The lexicon based approach inspect the message vocabulary in its dictionaries however it is studied that it is less accurate and varies across domains (Dhaoui, Webster, Tan 2017). Alternatively, Machine learning is more accurate, however the biggest drawback is that the training dataset will have to be manually classified beforehand and involves deeper pre-processing and training stage. Dhaoui, Webster, Tan (2017) & Wicaksono et al (2014) have conducted a research that ensemble approach of these 2 techniques will improve the model and are able in generating automatic corpus/training data.

3.2. Proposed Approach

The proposed method is illustrated in figure 3.2.1. is adapted from various journal article. Firstly, we'll collect our data from publicly used frequently such as Twitter, Facebook, Instagram. APIS, parsing on keywords on topic such as iphone, trump, and geotag (optional) is an example of this use (Stieglitz et al 2018). Following, Anwar et al (2015) we'll filter the data by semantic relation with Stanford Type Dependencies

Manual (SNLP) to exclude meaningless messages. It has been argued that there's 50 dependencies that convey information while others does not aid the results (Anwar et al 2015). Spam detection, fake news will also be excluded and then stored to the database.

Afterwards, we'll remove URL, username and timestamp, create token of words and tried to map Emoji, Abbreviation, metaphor to their textual meaning using dictionaries (Padtokar V & I 2016). The problem however is that allowing the machine to identify what's a metaphor or literal expression (Sulis et al 2016) which haven't been fully research.

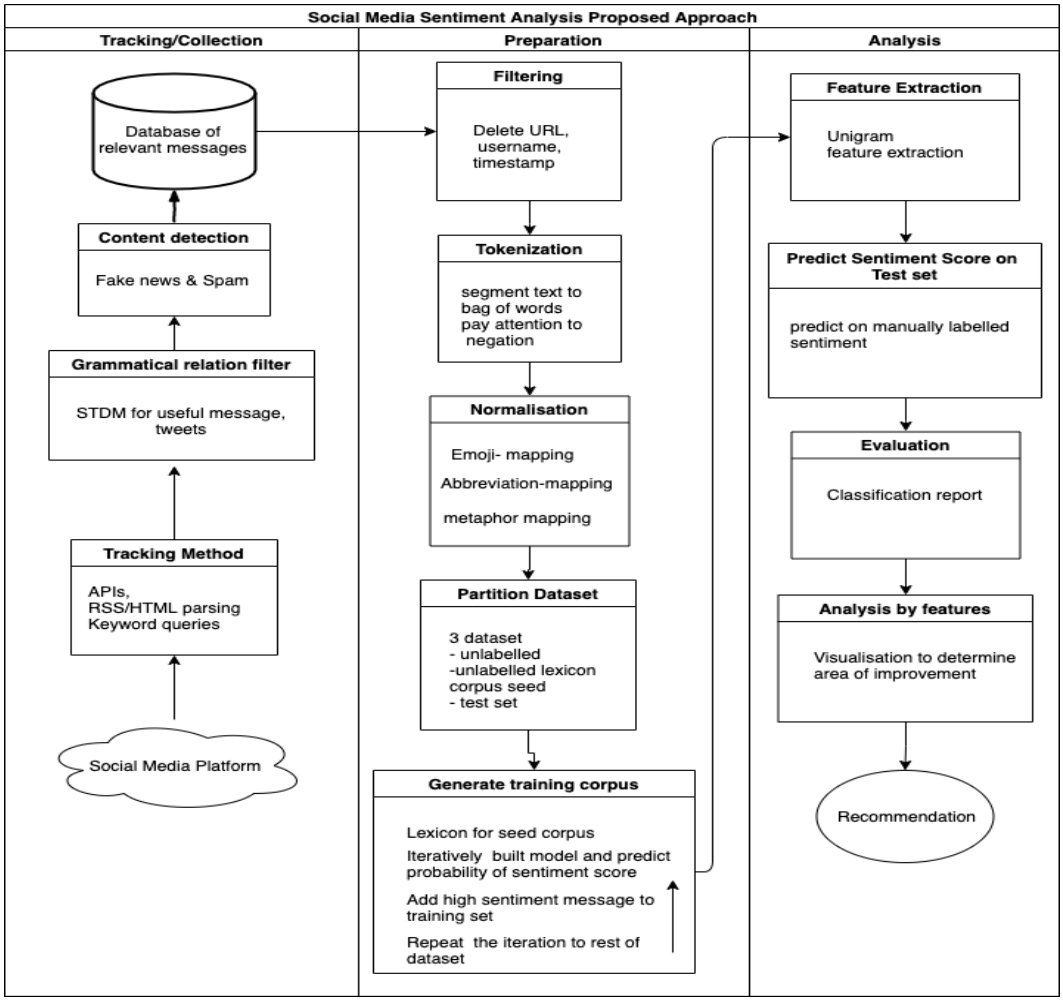

Figure 3.2.1. Proposed Approach Flow chart(original 2019)

As the combined approach was research to be better, we'll first generate a training corpus using lexicon approach, as manual labelling is time consuming. As demonstrated by Wicaksono et al (2014), we'll have the test set manually label by

social expert, as they're reliable in classifying sentiment, and have unlabelled data set, with label training set by lexicon on its specific domain. Their approach, is to iteratively use machine learning (Max entropy) to classify and give probability of unlabelled sets so that they can add new training instances by keeping the one with high probability as they are likely to be accurate. Reason being is that ME is more accurate then lexicon and works well in expanding training data (Wicaksono et al 2014).

Unigram feature extraction is then used as they provide more coverage and unlikely influenced by sparsity issue(Wicaksono et al 2014). Their experiments also concluded that lexicon is superior than Clustering sentiment labelling, and using Max Entropy it produce 80& F1, on the manually built dataset (Wicaksono et al 2014).

Model	Prec(%)	Rec(%)	F1(%)
BASELINE			
Naive Bayes	75.47	58.98	66.21
Maxent	78.36	74.85	76.56
LEX-METHOD			
Naive Bayes	76.24	64.37	**69.80**
Maxent	81.90	79.94	**80.91**
CLS-METHOD			
Naive Bayes	73.11	46.40	56.77
Maxent	80.00	63.47	70.78

Figure 3.2.2. Classifier Evaluation on proposed method(Wicaksono et al 2014)

To determine the issue of the overall sentiments inputting the average sentiment score based on features will uncover the problem. For instance, iPhone 6 have negative score on touch, as the phone are easily bent (Anwar et al 2015). Another example is He et al (2015) where he compared the score on features on Obama & Romney election, and find out that job and tax is the biggest and least issues for the voting. As we managed to uncover the cause of sentiments, it will give notice of what area to be improved in order to change people opinion/support.

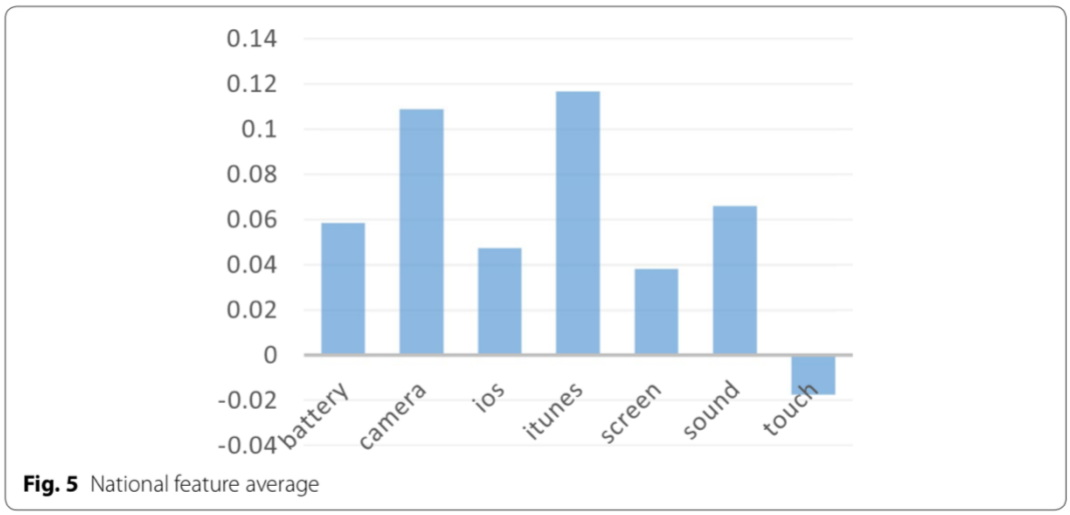

Figure 3.2.3. Analysis on Sentiment score based on feature (Anwar et al 2015)

4. Ethical Consideration

Social Media studies have become the centre of privacy concerns, Although, post and messages is design to show or give personal opinion of users to public, some groups and accounts are set to private (Michaelidou & Michevski 2019). Some users are not even aware of companies analysing their information. The social implication of this according to Michaelidou & Michevski (2019), is that users might be more paranoid and deliberately give false information, protests as they don't want to disclose information. Further policy and regulation, consent matter need to be addressed to lower the risk of unethical research as well as employing benefits of SMA.

5. References

Anwar Hridoy, S., Ekram, M., Islam, M., Ahmed, F. & Rahman, R. 2015,' Localized twitter opinion mining using sentiment analysis', *Decision Analytics*, vol.2, no.1.

Dhaoui, C., Webster, C. & Tan, L. 2017, 'Social media sentiment analysis: lexicon versus machine learning', *Journal of Consumer Marketing*, vol.34, no.6, pp.480-488.

Gomes, R. & Casais, B. 2018, 'Feelings generated by threat appeals in social marketing: text and emoji analysis of user reactions to anorexia nervosa campaigns in social media', *International Review on Public and Nonprofit Marketing*, vol.15, no.4, pp.591-607.

Hammer, M. 2017. ‘Ethical Considerations When Using Social Media for Research’, *Oncology Nursing Forum*, vol.44, no.4, pp.410-412.

He, W., Wu, H., Yan, G., Akula, V. & Shen, J. 2015, ‘ A novel social media competitive analytics framework with sentiment benchmarks’, *Information & Management*, vol.52, no.7, pp.801-812.

Karami, A., Bennet L.S. & He, X.2018, ‘Mining Public Opinion about Economic Issues: Twitter and the U.S Presidential Election’, International Journal of Strategic Decision Sciences, vol.9, no.1, 18-28.

Michaelidou, N. & Micevski, M.2019. ‘Consumers' ethical perceptions of social media analytics practices: Risks, benefits and potential outcomes’,*Journal of Business Research*, vol.104, pp.576-586.

Patodkar, V. & I.R, S. 2016. ‘Twitter as a Corpus for Sentiment Analysis and Opinion Mining’. *IJARCCE*, vol.5, no.12, pp.320-322.

Stieglitz, S., Mirbabaie, M., Ross, B. & Neuberger, C. 2018, ‘Social media analytics – Challenges in topic discovery, data collection, and data preparation’ *International Journal of Information Management*, vol.39, pp.156-168.

Sulis, E., Irazú, D., Rosso, P., Patti, V. and Ruffo, G. (2016).’ Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not’, *Knowledge-Based Systems*, no.108, pp.132-143.

Wicaksono, A.F., Vania, C., Distiawan, B., & Adriarni, M. 2014, ‘Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets’, *Faculty of Computer Science University of Indonesia,* pp. 185-194.

Zhang, W., Xu, M. & Jiang, Q. 2018, ‘Opinion Mining and Sentiment Analysis in Social Media: Challenges and Applications’, *HCI in Business, Government, and Organizations*, pp.536-548.