RESEARCH ARTICLE OPEN ACCESS

MULTI-STRATEGY SENTIMENT ANALYSIS OF CONSUMER REVIEWS WITH PARTIAL PHRASE MATCHING

R.Navin Kumar M.C.A.,M.Phil.,[1] , S.Sneha[2]

[1]Assistant Professor, Department of Computer Applications, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.

[2]Final MCA, Department of Computer Applications, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.

Email: [1]navinsoccer07@gmail.com, [2]sneha775352@gmail.com

Abstract:
Sentiment analysis is useful in commercial intelligence application environment and recommender systems, because it is a very convenient channel for the two ends of the supply to communicate. In the sentiment analysis, many strategies and techniques were used, such as machine learning, polarity lexicons, natural language processing, and psychometric scales, which determine different types of sentiment analysis, such as assumptions made, method reveals, and validation dataset. Since Internet has become an excellent source of consumer reviews, the area of sentiment analysis (also called sentiment extraction, opinion mining, opinion extraction, and sentiment mining) has seen a large increase in academic interest over the last few years. Sentiment analysis mines opinions at word, sentence, and document levels, and gives sentiment polarities and strengths of articles. As known, the opinions of consumers are expressed in sentiment phrases. Traditional machine learning techniques can not represent the opinion of articles very well. This project proposes a multi-strategy sentiment analysis method with semantic similarity to solve the problem with partial phrase matching. Naïve Bayes classification is also applied to find the probability of data distribution in various category of data set. The project is designed using R Studio 1.0. The coding language used is R 3.4.4.

Keywords: Sentiment Analysis, Naïve Bayes Classification, Multiple Strategy, Machine Learning.

I. INTRODUCTION

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered. So what should a brand do to capture that low hanging fruit?

With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. We believe it is important to classify incoming customer conversation about a brand based on following lines:

1. Key aspects of a brand's product and service that customers care about.
2. Users' underlying intentions and reactions concerning those aspects.

These basic concepts when used in combination become a very important tool for analyzing millions of brand conversations with human level accuracy. Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral. Intent analysis steps up the game by analyzing the user's intention behind a message and identifying whether it relates an opinion, news, marketing, complaint, suggestion, appreciation or query.

The Internet is currently not only an important source of information, but also a platform of expressing views and sharing experiences. In this network, we can easily collect reviews about products or services. Sentiment analysis is useful in commercial intelligence application environment and recommender systems [1], [2], because it is a very convenient channel for the two ends of the supply to communicate. In the sentiment analysis, many strategies and techniques were used, such as machine learning [3], polarity lexicons [4], natural language processing, and psychometric scales, which determine different types of sentiment analysis, such as assumptions made, method reveals, and validation datasets.

Currently, sentiment analysis is made at three consecutive levels: a) word, b) sentence, and c) document, in which sentence and document are usually used in most previous studies. The word, the fundamental, and

consequently the more significant and more challenging level, however, are never studied. chinese language actually uses short sentiment phrases for one / two Chinese characters which are the most fuzzy in meaning. Traditional machine learning techniques cannot denote this characteristic. Hence, a hybrid sentiment analysis is studied here in this study; which comprehensively studies a) Zadeh's fuzzy set theory, b) machine learning theory, and c) the method based on polarity lexicons. It also considers adversative conjunctions, such as i) "(but)", ii) "(while)", iii) "(however)" etc.

Western scholars proposed the sentiment analysis earlier. They determined first the tendency (sentiment) of words / phrases and quantified them as the measure of real values, [6] that can be further utilized to find sentiment tendency of both sentences as well as paragraphs. They analyzed the sentiment tendency

The three standard machine learning algorithms for sentiment analysis are a) NB (Naive Bayes), b) ME (MaxEnt, or Maximum Entropy), and c) SVMs (Support Vector Machines). For simplicity of experiment, the authors here only choose NB and SVMs.

II. LITERATURE REVIEW

In the paper [1], the authors stated that the feasibility of automatic recognition of recommendations is supported by empirical results. First, Usenet messages are a significant source of recommendations of Web resources: 23% of Usenet messages mention Web resources, and 30% of these mentions are recommendations. Second, recommendation instances can be machine-recognized with nearly 90% accuracy. Third, some resources are recommended by more than one person. These multi-confirmed recommendations appear to be significant resources for the relevant community.

Finally, the number of distinct recommenders of a resource is a plausible measure of resource quality. A comparison of recommended resources with resources in FAQs(lists of Frequently Asked Questions maintained by human topic experts) indicates the more distinct recommenders a resource has, the more likely it is to appear in the FAQs. PHOAKS is distinguished from other recommender systems by two major design principles: role specialization and reuse.

Many recommender systems, particularly ratings-based systems [11, 13, 14], are builton the assumption of role uniformity. They expect all users to do the same types of work in return for the same types of benefits. In the case of ratings-based systems, for example, everyone rates objects of interest. Yet there is evidence that people naturally prefer to play distinct producer/consumer roles in the information ecology [12]; in particular, only a minority of people expend the effort of judging information and volunteering their opinions to others. Independently, they have observed such role specialization in Netnews; authors volunteer long lists of recommended Web resources at a stable, but low rate. PHOAKS assumes the roles of recommendation provider and recommendation recipient are specialized and different.

What Counts as a Recommendation?

The basic idea of collaborative filtering is people recommending items to one another. Readers of Usenet news know this is a normal practice in newsgroups. Posters often volunteer their impressions and opinions about all sorts of items, including Web pages. They may state what a page is useful for and how useful it is: PHOAKS searches messages for mentions of Web pages (URLs) and counts a mention as a recommendation if it passes a number of tests.

First, the message must not be cross-posted to too many newsgroups. Messages posted to a large number of groups are so general they are not likely to be thematically close to any of the groups. Second, if the URL is part of a poster's signature or signature file, it is not counted as a recommendation. Third, if the URL occurs in a quoted section of a previous message, it is ruled out. Fourth, if the textual context surrounding the URL contains word markers that indicate it is being recommended and does not contain makers that indicate it is being advertised or announced, then it is categorized as a recommendation. They have developed rather complicated categorization rules that implement this basic strategy to distinguish the different purposes for which Web resources are mentioned.

They said that their future work as follows:

First, they are continuing to analyze the relationship between resources recommended in Usenet messages and resources that appear in FAQs. They are especially interested in the temporal dimension. So, for example, they will determine to what extent Usenet messages are a leading indicator of FAQ content. Second, they will use FAQs to enhance their interface to recommendation data. For example, one will be able to go from a resource to references to that resource in FAQs. Thus, we intend to combine the best of automatically mined recommendations (e.g., timeliness) with human-determined recommendations (e.g., long-term relevance and quality). Third, they will apply their generic filtering architecture to extract other types of information from electronic messages.

In this paper [2], the author stated that a collaborative exploration system is being proposed that helps users to explore recommendations from various viewpoints. Given ratings and reviews on movies from reviewers, the system provides "virtual reviewers" that represent particular viewpoints. Each virtual reviewer navigates the user by recommending and characterizing both movies and reviewers according to its viewpoint. The author has developed a browsing method with virtual reviewers and visual interfaces. Collaborative filtering is an information retrieval technique that utilizes knowledge from other users [15] [16]. It can deal with a user's subjective "taste" for items such as movies and music based on users' ratings.

However, the filter will not automate all of the user's tasks to obtain information from other users: the user needs a query interface to access items actively. The user sometimes has more specific (and often temporary) needs than her or his general interests on which filtering results are based. In such cases, the user will need to explore for items that meet the specific needs. The user will also need to explore for less predictably but potentially interesting items the filter might exclude, that is, serendipitous information.

The author proposed a collaborative exploration system that generalizes the automatic recommendation technique of collaborative filtering in order to help users to explore recommendations from various viewpoints. Information about a certain kind of items can be obtained by consulting people who know those items well. Collaborative information exploration virtualizes this process by using rating data. The author has developed a movie database that realizes users' collaborative exploration of movie reviews and ratings given by a number of reviewers on the Internet. Based on the rating data, the system provides a "virtual reviewer" that has a particular viewpoint. A viewpoint is represented as a set of movies, and a virtual reviewer simulates a reviewer who likes these movies. A virtual reviewer navigates the user by recommending and characterizing both movies and reviewers according to its viewpoint.

In this paper, the author described a virtual reviewer and its functions. The author then proposed a browsing method that uses virtual reviewers and an automatic clustering technique. The author also introduced a visual embodiment of a virtual reviewer to realize visual explanation and querying of movies and reviewers from various viewpoints.

COLLABORATIVE FILTERING AND BROWSING

The Tapestry system [17], which coined the term "collaborative filtering," is a mail system that filters mail or news articles based on annotations given by other users. The system supports a query language TQL that enables the user to find articles that meet specific needs. The user, for instance, can obtain articles recommended by a specified person. The system, however, is not suitable for a large community in which users don't necessarily know each other: it is meant to support a work group in which the user can specify appropriate actual users to get information. Recent popular collaborative filtering systems [15] [16] automate selection of users who share interests with the user. By computing similarity of the users based on their rating patterns, the system provides the user with similar users (neighbors) and items they recommend. However, as described in the previous section, the user needs tools for exploration in addition to automatic filters.

Collaborative browsing has been studied to help multiple users to collaborate in browsing synchronously or asynchronously. These systems provide a group of users with tools to share processes or histories of browsing. For a large community, however, the systems will need additional ways to organize diverse information from users with various interests or tastes. Let's Browse helps a group of users to find items of common interest even if they don't know each other. The system, designed for real-time browsing sessions, provides appropriate topics for specific participants but does not search appropriate reviewers for specific interests. Their system is designed for exploration of various users (reviewers) in a large community as well as exploration of items (movies). The user can automatically or manually create various virtual reviewers, each of which consists of a set of movies and a set of users who like those movies. While a filter uses a single durable profile to handle the user's general interests, multiple virtual reviewers provide various viewpoints to help the user to explore both reviewers and movies.

In this paper [3], the authors stated that Microblogging, like Twitter1, has become a popular platform of human expressions, through which users can easily produce content on breaking news, public events, or products. The massive amount of microblogging data is a useful and timely source that carries mass sentiment and opinions on various topics.

Existing sentiment analysis approaches often assume that texts are independent and identically distributed (i.i.d.), usually focusing on building a sophisticated feature space to handle noisy and short messages, without taking advantage of the fact that the microblogs are networked data. Inspired by the social sciences findings that sentiment consistency and emotional contagion are observed in social networks, they investigated whether social relations can help sentiment analysis by proposing a Sociological Approach to handling Noisy and short Texts (SANT) for sentiment classification.

In particular, they presented a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the supervised learning process; and utilize sparse learning to tackle noisy texts in microblogging. An empirical study of two real-world Twitter datasets shows the superior performance of our framework in handling noisy and short tweets.

Microblogging services are extensively used to share information or opinions in various domains. With the growing availability of such an opinion-rich resource, it attracts much attention from those who seek to understand the opinions of individuals, or to gauge aggregated sentiment of mass populations. For example, advertisers may want to target users who are enthusiastic about a brand or a product in order to launch a successful social media campaign. Aid agencies from around the world would like to monitor sentiment evolutions before, during, and after crisis to assist recovery and provide disaster relief.

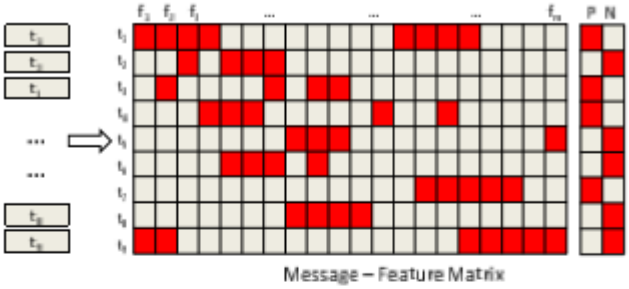
The sheer volumes of microblogging data present opportunities and challenges for sentiment analysis of these noisy and short texts. Sentiment analysis has been extensively studied for product and movie reviews, which

differ substantially from microblogging data. Unlike standard texts with many words that help gather sufficient statistics, the texts in microblogging only consist of a few phrases or 1-2 sentences.

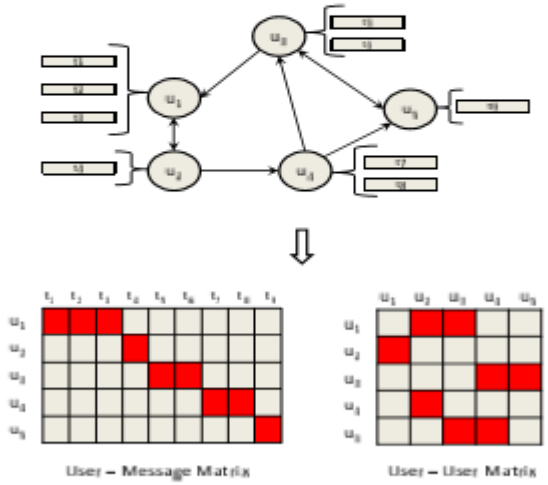
Also, when composing a microblogging message, users may use or coin new abbreviations or acronyms that seldom appear in conventional text documents. For example, messages like “It is coooooool”, “OMG :-(”, are intuitive and popular in microblogging, but some are not formal words. It is difficult for machines to accurately identify the semantic meanings of these messages, though they provide convenience in quick and instant communications for human beings. Existing methods rely on pre-defined sentiment vocabularies [9], which are highly domain-specific.

Meanwhile, microblogging platforms often provide additional information other than text. For example, in Figure 1, we depict two kinds of data available in microblogging. Figure 1(a) shows the content of messages, in the form of a message-feature matrix. Traditional methods measure the similarity between text documents (messages) purely based on content information.

A distinct feature of microblogging messages is that they are potentially networked through user connections, which may contain useful semantic clues that are not available in purely text-based methods. Besides content information, relations between messages can be represented via a user-message matrix and a user-user interaction matrix, as shown in Figure 1(b). Traditional methods, if applied directly to the microblogging data, do not utilize the social relation information. In social sciences, it is well-established that emotions and sentiments play a distinct role in our social life and correlate with our social connections. When experiencing emotions, people do not generally keep the emotions to themselves, but rather, they tend to show them. Also, people tend to“catch”others’ emotions as a consequence of facial, vocal, and postural feedback, which has been recognized as emotional contagion in social sciences.



(a) Data Representation of Message Content



(b) Data Representation of Social Relations

Figure 1: Data Representation of Text and Social Relation Information in Microblogging

Emotional contagion may be important in personal relationships because “it fosters behavioral synchrony and the tracking of the feelings of others moment-to-moment even when individuals are not explicitly attending to this information”. As a consequence of emotional contagion, Fowler and Christakis [10] reported the spread of happiness in a social network.

Two social processes, selection and influence, are proposed to explain the phenomenon: people befriend others who are similar to them, or they become more similar to their friends over time (Social Influence). Both explanations suggest that connected individuals are more likely to have similar behaviors or hold similar opinions. Inspired by this sociological observation, we explore the utilization of social relation information to facilitate sentiment analysis in the context of microblogging

In this paper, they aimed to provide a supervised approach to sentiment analysis in microblogging by taking advantage of social relation information in tackling the noisy nature of the messages. In particular, they first investigated whether the social theories exist in microblogging data. Then they discussed how the social relations could be modeled and utilized for supervised sentiment analysis.

Finally, they conducted extensive experiments to verify the proposed model. The main contributions of this paper are as follows:

• They formally define the problem of sentiment analysis in microblogging to enable the utilization of social relations for sentiment analysis;

• By verifying the existence of two social theories in microblogging, they built sentiment relations between messages via social relations;

• They presented a novel supervised method to tackle the noisy and short texts by integrating sentiment relations between the texts; and

• They empirically evaluate the proposed SANT framework on real-world Twitter datasets and elaborate the effects of social relationships on sentiment analysis.

They concluded that different from texts in traditional media, microblogging texts are noisy, short, and embedded with social relations, which presents challenges to sentiment analysis. In this paper, they proposed a novel sociological approach (SANT) to handle networked texts in microblogging. In particular, they extracted sentiment relations between tweets based on social theories, and model the relations using graph Laplacian, which is employed as a regularization to a sparse formulation. Thus the proposed method can utilize sentiment relations between messages to facilitate sentiment classification and effectively handle noisy Twitter data. We further develop an optimization algorithm for SANT.

Experimental results show that the user-centric social relations are helpful for sentiment classification of microblogging messages. Empirical evaluations demonstrate that their framework significantly outperforms the representative sentiment classification methods on two real-world datasets, and SANT achieves consistent performance for different sizes of training data, a useful feature for sentiment classification.

This work suggests some interesting directions for future work. For example, it would be interesting to investigate the contributions of different sentiment relations to sentiment classification. Other information, like spatial-temporal patterns, could be potentially useful to measure the sentiment consistency of people as well. For example, people in Miami might be happier about the temperature than people in Chicago during winter time. They could further explore how sentiments diffuse in the social network and how people's sentiments correlate with internal (their friends) and external (public events) factors. With the analysis, it is possible for them to understand the differences of sentiment between the online world and physical world.

III. PROPOSED METHODOLOGY

In existing system, data set is taken as records from Excel worksheet with category in second column. Preprocessing work such as stop word removal, stemming and Unicode removal are being done. Punctuation marks are removed. All characters are converted in lower case.

Then two words, three words combinations are found out. If the count is above the given threshold among all the records, then the words are treated as valid phrases. These phrases conditional probability is found out among all categories which become Naïve Bayes Classification work. In addition, synonym words and phrases are kept in separate file.

Before taken for probability finding, the phrases are replaced with synonym words so that two different phrases become same in semantic similarity/fuzziness concept.

IV. FINDINGS

- Exact phrase match is taken for conditional probability. That is two phrases in two records must match exactly during probability finding.
- Skip grams (elimination of a word in phrase) is not checked during classification.
- Partial phrase is not checked during classification.
- In proposed system, like existing system, data set is taken as records from Excel worksheet with category in second column. Preprocessing work is carried out. Then words combinations are found out and valid phrases are gathered.
- These phrases conditional probability is found out among all categories which become Naïve Bayes Classification work. In addition, synonym words replacement is also made. Moreover, partial phrases like two words in one sentence and three words in other sentence are also treated as same phrases during naïve bayes classification.
- Exact phrase match is not taken for conditional probability. That is two phrases in two records need not match exactly during probability finding.
- Skip grams (elimination of a word in phrase) is checked during classification.
- Partial phrase is checked during classification.

V. CONCLUSION

A new system for the computation of oppositeness and strengths of sentiment expressions is proposed in this design, which could be used to dissect semantic similarity of rulings indeed with partial expression matching. It uses a probability value, rather than a standard value for the opposition strengths of sentiment expressions, compared with the conventional styles. According to the oppositeness and strengths of those expressions, it proposesmulti-strategy sentiment analysis system independently grounded on NB. Particularly, in the system grounded on NB, it considers adversative convergences. The system could be used for the sentiment analysis of documents. The feasibility and effectiveness of the system is proved. In future, this design may concentrate on how the similarity

could be plant out using emoticons and Unicode character representations.

REFERENCES

[1] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter, ‘‘PHOAKS: A system for sharing recommendations,’’ Commun. ACM, vol. 40, no. 3, pp. 59–62, 1997.

[2] J. Tatemura, ‘‘Virtual reviewers for collaborative exploration of movie reviews,’’inProc.5thInt.Conf.Intell.UserInterfaces,2000,pp. 272–275.

[3] X. Hu, L. Tang, J. Tang, and H. Liu, ‘‘Exploiting social relations for sentiment analysis in microblogging,’’ in Proc. 6th ACM Int. Conf. Web Search Data Mining, 2013, pp. 537–546.

[4] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, ‘‘Opinion extraction, summarizationandtrackinginnewsandBlogCorpora,’’inProc .AAAISpringSymp., Comput. Approaches Anal. Weblogs, 2006, pp. 100–107.

[5] A. C. König and E. Brill, ‘‘Reducing the human overhead in text categorization,’’ in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 598–603.

[6] F. Zhou, R. J. Jiao, and J. S. Linsey, ‘‘Latent customer needs elicitation by use case analogical reasoning from sentiment analysis of online product reviews,’’ J. Mech. Des., vol. 137, no. 7, p. 071401, 2015

[7] E. Cambria, ‘‘Affective computing and sentiment analysis,’’ IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016

[8] M. Govindarajan, ‘‘Sentiment analysis of restaurant reviews using hybrid classification method,’’ in Proc. 2nd IRF Int. Conf., Chennai India, Feb. 2014, pp. 127–133.

11. Hill, W.C., Stead, L., Rosenstein, M., and Furnas, G. Recommending and evaluating choices in a virtual community of use. In Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI’95 (Denver, May 7–11). ACM Press, New York, N.Y., pp. 194–201.

12. Maltz, D., and Ehrlich, K. Pointing the way: Active collaborative filtering. In Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI’95 (Denver, May 7–11). ACM Press, New York., N.Y., pp. 202–209.

13. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of netnews. In Proceeding of the ACM 1994 Conference on Computer Supported Cooperative Work. pp. 175–186.

14. Shardanand, U., and Maes, P. Social information filtering: Algorithms for automating “word of mouth.” In Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI’95(Denver, May 7–11). ACM Press, New York. N.Y., pp. 210–217.

15. Resnick, P., et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proceedings of ACM CSCW’94, 1994, 175-186.

16. Shardanand, U. and Maes, P. Social Information Filtering: Algorithms for Automating “Word of Mouth.” Proceedings of ACM CHI’95, 1995, 210-217.

17. Goldberg, D., et al. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM, vol. 35, No. 12, 1992, 61-70.

[18] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39:165–210, 2005.

[19] J. Fowler and N. Christakis. The dynamic spread of happiness in a large social network. BMJ: British medical journal, 2008.

Wiebe, J., Wilson, T., and Bell, M. Identify collocations for recognizing opinions. Proceedings of ACL/EACL2001 Workshop on Collocation.2001.

Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on EMNLP,pages 79-86. 2002.

Dave, K., Lawrence, S., and Pennock, D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. WWW 2003,pages 519-528, 2003.

Riloff, E. and Wiebe, J. Learning extraction patterns for subjective expressions. Proceedings of the 2003 Conference on EMNLP,pages 105-112. 2003.

Liu, B., Hu, M. and Cheng, J. Opinion Observer: Analyzing and comparing opinions on the web. WWW 2005, pages 342-351. 2005.

Cardie, C., Wiebe, J., Wilson, T. and Litman, D. Combining lowlevel and summary representations of opinions for multiperspective question answering. Proceedings of AAAI Spring Symposium Workshop,pages 20-27. 2004.

Takamura, H., Inui, T. and Okumura, M.. Extracting Semantic Orientations of Words Using Spin Model. ACL 2005, pages 133-140. 2005.

Adomavicius, G. ,Tuzhilin, A. (2005), "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 6.