SENTIMENT ANALYSIS OF GOVERNMENT SOCIAL MEDIA TOWARDS AN AUTOMATED CONTENT ANALYSIS USING SEMANTIC ROLE LABELING

Siti Salwa Hasbullah¹ and Rita Zaharah Wan-Chik²

¹Malaysian Institute of Information Technology, Universiti of Kuala Lumpur (UniKL)
sitisalwah@unikl.edu.my

²Centre for Research and Innovation, Universiti of Kuala Lumpur (UniKL)
ritazaharah@unikl.edu.my

ABSTRACT

In this paper, we propose to develop an automated content analysis tool to help the Malaysian government’s cyber and legal advisors, as well as the government leaders, to understand public sentiment via their comments which are posted on the official government leaders’ or ministerial social media sites (i.e., Twitter, Facebook, etc.). In this study, we explore and apply the Semantic Role Labeling (SRL) techniques that generate new methods to filter and classify the social media content data set, advancing the state of the art in sentiment detection approaches. This proposed automated content analysis tool would be able to provide a platform to measure the impact of public sentiment over the government leader’s postings, and the public’s comments, on their officials’ social media sites. The results and findings from the impact measurement could then be used as a recommendation in the developing or reviewing the national’s cyber communication policy.

Keywords: Sentiment Analysis, Natural Language Processing, Semantic Role Labeling, Social Media, Automated Content Analysis

1. Introduction

The Malaysian government has realized that the advent of social media is a great vehicle to get closer to the citizens. It can provide deep insights into the needs of the citizens and thus could help the government to move closer to the citizen, where the priorities and services would be driven according to citizen needs rather than government’s capability. The Malaysian government’s leaders and ministerial offices have now their own official Facebook and Twitter accounts to get connected with the citizens and to help the government to understand its citizens closely. This research will investigate the many responses to these official social media accounts, focusing on the emotions expressed, whether it be people are making suggestions, complaining or just passing information. The main objective of this research is to design and develop new methods and techniques to classify polarity and sentiment expressions in the social media content. It aims to be able to better assist the governments on further action plans and strategies in public communications. A framework and recommendation for the government’s public communicative actions based on this sentiment polarity and classification will be developed.

By using a state-of-the-art base-line technique called Semantic Role Labeling (SRL), this study will work by labeling the text using phrase structure and syntactic parsing via a context-free or context-sensitive grammar, a dependency parser, or other related method. It will then be integrated with General Architecture for Text Engineering (GATE) proposed by Cunningham et al. (2002), which is an...
architecture that contains functionality for plugging in all kinds of NLP software, such as POS taggers, sentence splitters and Named Entity recognizers. This technique has been used in a wide variety of text processing research and development projects. (Diana and Mar, 2014).

This study will focus at emotion detection, specifically on the emotions re-expressed or any comments made to classify the comment into either positive or negative comments. The SRL technique will be used in the process of annotating the predicate-argument structure in text (i.e.; comments) with semantic labels. This is done by labeling the text using phrase structure and syntactic parsing via a context-free or context-sensitive grammar, a dependency parser, or other related method. This process would generally be performed in an earlier stage and used as input to sentiment analysis. This technique is believed to be most relevant in identifying the argument in syntactic parsing, especially, in the very first stage to express sentiment analysis (Lu and Tsou, 2010). The evaluation metrics will then be conducted to measure the sentiment emotion detection accuracy by using Precision (Pre) and Recall (Rec) as the weightage process by (Paolo and Gabriele, 2008).

2. Related Work

There has been an explosive growth of social application tools, and while these social media have been widely adopted publicly, organizations are only recently realizing their potential. In order to relate with recent citizen views over Malaysian government leaders, there has been increasing interest in constructing general-purpose public opinion classifiers because of their potential on public opinion analysis by Pang and Lee (2008). Solutions to these issues are sought by research in natural language processing (NLP), the artificial intelligence (AI) discipline that deals with the automatic treatment of natural language in text or speech. Alongside these traditional natural language processing areas, recent years have seen the birth and constant growth in importance of two new research tasks, subjectivity and sentiment analysis by Ruppenhofer et al. (2012). Together with emotion detection, a task that expands beyond the field of natural language processing, subjectivity and sentiment analysis are parts of the wider field in Artificial Intelligence by Abidin et al. (2011).

The proposed method of SRL, also called shallow semantic parsing, is a popular semantic analysis technique and has become a focus of research in computational linguistics during the recent years (Liu et al., 2010; Erk and Pado, 2005). By using this technique, we can detect the semantic arguments associated with the predicate or verb of a sentence and perform their classification into their specific opinion analysis. A previous study by Liu et al. (2010) have used this technique to prove their method in sentiment analysis research which investigated on a Twitter-based criminal incident prediction using a state-of-the-art NLP to extract the semantic contents of the tweets. They used SRL on semantic understanding of events mentioned in tweets, the entities involved in the events and the roles of the entities with respect to the events. They used Latent Dirichlet Allocation (LDA) model to extract the topics in Twitter, and applied Generalized Liner Model (GLM) model for evaluation prediction using new tweets, Wang et al, (2012). The use of LDA/GLM model was able to predict future hit-and-run incidents, although there was a limited amount of testing data.

Another study by Liu et al. (2011) is one where they used a clustering method that gathered similar tweets that were then grouped together. They proposed two-stage labelling; one labeller conducts semantic role labelling to get statistical information, such as the predicate or argument or role triples that occur frequently, from its highly confidently labelled results; then in the second stage, another labeller performs semantic role labelling with such statistical information to refine the
results. Experimental results on a human annotated dataset show that their technique remarkably improves the precision performance.

There is a growing research interest in NLP with researches ranging from document level classification by Pang and Lee, 2008 to learning the polarity of words and phrases by Esuli and Sebastiani, (2006). With the character limitation on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis according to Yu et al. (2003) and Kim et al. (2004). However, the informal and specialized language used in tweets, as well as the very nature of the microblogging domain make Twitter sentiment analysis a very different task. In recent years, there have been a number of studies looking at Twitter sentiment (Jansen et al. (2009); Pak and Paraozek (2010); O’Connor et al. (2010); Tumasjan et al. (2010); Bifet and Frank, (2010); Barbosa and Feng, (2010); and Davidov et al. (2010). Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data (Pak and Paraozek (2010); Bifet and Frank, (2010); and Barbosa and Feng, (2010)) which exploit existing Twitter sentiment sites for collecting training data. The research by Davidov et al. (2010) also uses hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 3-way polarity classification. Diana and Mark (2014) studied the impact of sarcasm tweets and have developed a hashtag tokenizer for GATE so that sentiment and sarcasm found within hashtags can be detected more easily.

Figure 1 : Screenshot of retokenized hashtags in GATE

An online tool was introduced by Duc et al. (2013) that proposed a fuzzy propagation modelling for opinion mining by sentiment analysis of online social networks. A practical system, called TweetScope, was developed to efficiently collect and analyse all possible tweets from customers. Their TweetScope application is able to collect tweets from Twitter efficiently by using Twitter Stream service; the user can access and generate useful data that represents how information is propagated on Twitter, using the definition model with the collected data. However, there is a lacking of the number of tweets and scale of friendship relationship network and TweetScope
cannot work online with dynamic information propagation where it has to fetch all data to a local database before processing and analysing, hence requiring more on storage memory and performance ability of the computer system. The research found that tweets containing emotional words are most frequently retweeted. The values of their linguistic variables are concentrated around the coordinate origin point in space of information propagation linguistic variables, which represent high frequency of retweeting. However, there are many reasons why a user retweets a tweet; hence the experiential result does not mention that non-emotional tweets are not more retweeted. Applying the model to a large scale dataset, analysing and visualization help the user understand clearly how information is diffused and how to increase the effect of the information propagation using emotional words in the advertising content on marketing and business field. Moreover, mash-up applications adopted by Jung, (2012) can be implemented by using external open APIs.

Social media analytics is the practice of gathering data from blogs and social media websites and analysing that data to make business decisions. The most common use of social media analytics is to mine customer sentiment. Andranik et al. (2010) has shown in their research that the Twitter mirrors the aggregate public sentiment in terms of identifying things like stock market trends, citizen happiness, and candidate ideology. Twitter and Facebook are the most used social media services by the government by Durant and Smith, 2006.

The studies of Malaysian government social media posts are important to reach their social media activities, providing the most appropriate services that government give from government-to-user communication. By using this, governments should keep an eye on upcoming public opinion and use those that show how their citizens respond to the services (Mainka et al. (2014)).

Recently in Malaysia, the government leaders recognize the advantages of using social media to interact with the citizens. According to the present Malaysian Prime Minister, Dato’ Seri Najib Tun Razak’s speech, "Use of social media is needed in this time and age, however, we need to be careful because it could be leaked out in blogs without one realizing it," it can be clarified that Malaysians have to face the challenge of negative information overload on the Internet.

Malaysia guarantees internet freedom under the Multimedia Science Corridor plans, where the government has vowed not to filter the internet. However, the country has several acts on cyber laws which have been passed in the Parliament that according to Saripan et al. (2011) that including the Communications and Multimedia Act 1998, Communications and Multimedia Commission Act 1998, Digital Signature Act 1997 (Act 562), Computer Crimes Act 1997 and Telemedical Act 1997. This study is suggesting that there should be a tool or system that could help the Government to monitor the communication in the Government’s official social media sites. Currently, we still lack of an automated content analysis tools to assist the Government in the monitoring activities. Thus, this study is recommending developing an automated content analysis tool for sentiment analysis using sentiment role labelling.
3. Proposed Model

In this paper we propose a model that would help us to find a solution for sentiment analysis detection. This study will be looking at the process of text processing using SRL techniques together with rule-based model generating for content analysis detection. The following Figure 2 illustrates the conceptual model for this study that shows the steps that will be taken during the research. Each step will be explained in detail in the following paragraph.

![Conceptual Model for this research](image)

Figure 2 : Conceptual Model for this research

3.1 Analyze the current sentiment and opinion using Semantic Role Labeling

There are two main social media medium that are being used by the Malaysian government leader; Facebook and Twitter. These two social media medium help the top government leader to monitor and receive feedback and comment from the public of their services. As normal process, Facebook and Twitter will release all comments for every post and tweets send by the official government leader social media. Comments are extracted to perform the quantification of sentiments toward a particular issue. The opinion expressions will be classified for sentiment detection purposes. The process is started by extracting the public comments using a chosen crawler and stored into a database. The pre-processing task using SRL will produce processes such as sentence splitting, tokenization and part of speech tagging for every extracted comment. The classification process is to classify the subjectivity of the sentences which is focusing on determining subjective words and texts that mark the presence of opinions and evaluations, together with the objective words and texts, used to present factual information (Wiebe et al. 2005). The sentence sentiment classification is a way to analyze the subjective information in the text and then mine the opinion and content detection. SRL using syntactic clues solely is inadequate especially when the ambiguity phrases exist within the sentence or data testing.

3.2 The proposed rule-based model

The use of rule-based model is seen as a possible solution to improve the accuracy of the result. It is used as a way to store and manipulate knowledge to interpret information in a useful way. It is often used in AI applications and researches by Johansson (2008). It is particularly used for a method that often rapidly leads to a solution that is usually reasonably close to the best possible answer (Jurafsky, 2000)). In order to develop the rule-based model, a collection of existing rule-based model from the previous research is reviewed. Potential new rules are then developed based on a training set. The focus of research at this stage is to gather the existing rules which are based on...
natural language syntax to determine the elements of object oriented from natural language’s specification. In order to form new rules, a new training dataset which consists of all possible arguments of sentiments are used. The combination of the new and existing rules based such as rule base 1, rule based 2, rule based 3 and others will manually testing using the training dataset. The aim is to ensure that only suitable and optimized rule based will be chosen for the implementation of the sentiment detection.

3.3 Design an automated content analysis tool

An automated content analysis tool will be developed using JAVA language to perform the sentiment opinions, emotions and judgments detection. This proposed tool will be plugged-in with GATE, an integrated development environment for language processing components bundled with the most widely used information extraction system and a comprehensive set by Tablan et. al, (2014). GATE is an architecture that contains functionality for plugging in all kinds of natural language processing software, such as part-of-speech (POS) taggers, sentence splitters and named entity recognizers. GATE has been in development at the University of Sheffield since 1995 and has been used in a wide variety of research and development projects. Some research projects have been done using GATE-based sentiment analysis and opinion mining tools, specifically optimized for Twitter, blogs, comments, and other kinds of social media posts (Diana and Mark, (2014)). The use of GATE by Durant et al. (2006) helped the research in terms of text processing and architecture that contains functionality for plugging in all kinds of NLP software, such as part-of-speech (POS) taggers, sentence splitters and named entity recognizers. GATE as an architecture suggests that the elements of software systems that process natural language can usefully be broken down into various types of component, known as resources. Components are reusable software chunks with well-defined interfaces, and are a popular architectural form, used in Sun’s Java Beans and Microsoft’s .Net, for example. GATE general function includes components for diverse language processing tasks, e.g. parsers, morphology, tagging, Information.

3.4 Matrix Evaluation

In order to achieve more efficient and competitive automated content application tool, we assign weights from tested data set based on Precision (Pre) and Recall (Rec) matrix evaluation to increase the likelihood that it is correct. By using this evaluation matrix approach, this application tool as a whole will be useful for the government officials and the Malaysian agencies to help them monitor public opinion as an early warning system of possible disruption.

3.5 Proposed Policy and Recommendation

Based on the findings from this research, a framework and recommendation for the government’s public communicative actions based on this sentiment polarity and classification will be developed. This will help the government’s legal advisors such as the Royal Malaysian Police to control sedition among the public; the Federal government agency such as Malaysian Communication and Media Commission (MCMC) to control public’s access and communication to government’s official social media platform; the IT agency such as Malaysian CyberSecurity to help the government to create and sustain a safer cyberspace among its cyber-citizen.
4. Conclusion

Throughout research activities, there will be more new findings on the related research and these new research findings will help to improve the existing technology’s performance. By understanding the current research interest in the field of sentiment analysis, the idea for the content analysis based on Malaysian government leader’s social media is still rare and new. As mentioned, this research aims to study the Malaysian government’s leader official communication platform through social media such as Twitter and Facebook with the main objective is to design and develop an automated content analysis tool to assist for the sentiment analysis of public comments on the Government’s social media sites. The goal is to be able to better assist on governments’ further action plans and strategies on public communications. The main focus of this research is to classify polarity and sentiment expressions in the comments using the new rule based method that is integrated with the SRL techniques.

The implication from the findings of this study is would have a benefit to the society and especially for the Malaysian government Legal Advisors, Cyber Security Agencies and also to the Malaysian citizens as a whole. Current research in Malaysian opinion sentiment on social media is restricted more to looking inside the public and citizen comments without looking at different levels of people, especially at the top leaders in Malaysia. The collaboration works is ongoing with various government agencies to acquire the actual input and finding to this research context. The end product of this research, after proven to be successful, could be then be used by all government sectors and government-linked agencies on their social media platforms.
References


