Coupling Topic Modelling in Opinion Mining for Social Media Analysis

Xujuan Zhou
School of Information Systems,
University of Southern Queensland, Australia
xujuan.zhou@usq.edu.au

Xiaohui Tao
School of Agricultural, Computational and Environmental Sciences,
University of Southern Queensland, Australia
xtao@usq.edu.au

Md Mostafijur Rahman
School of Agricultural, Computational and Environmental Sciences,
University of Southern Queensland, Australia
Md.Rahman@usq.edu.au

Ji Zhang
School of Agricultural, Computational and Environmental Sciences,
University of Southern Queensland, Australia
ji.zhang@usq.edu.au

ABSTRACT

Many of social media platforms such as Facebook and Twitter make it easy for everyone to share their thoughts on literally anything. Topic and opinion detection in social media facilitates the identification of emerging societal trends, analysis of public reactions to policies and business products. In this paper, we proposed a new method that combines the opining mining and context-based topic modelling to analyse public opinions on social media data. Context based topic modelling is used to categorise data in groups and discover hidden communities in data group. The unwanted data group discovered by the topic model then will be discarded. A lexicon based opinion mining method will be applied to the remaining data groups to spot out the public sentiment about the entities. A set of Tweets data on Australian Federal Election 2010 was used in our experiments. Our experimental results demonstrate that, with the help of topic modelling, our social media analysis model is accurate and effective.

CCS CONCEPTS

• **Information systems → Collaborative and social computing systems and tools**; • **Computing methodologies → Machine learning approaches**;

KEYWORDS

Opinion mining, Topic modelling, Social media analysis, Online social networks

ACM Reference format:
Xujuan Zhou, Xiaohui Tao, Md Mostafijur Rahman, and Ji Zhang. 2017. Coupling Topic Modelling in Opinion Mining for Social Media Analysis. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
DOI: 10.1145/3106426.3106459

WI '17, Leipzig, Germany
© 2017 ACM. 978-1-4503-4951-2/17/08…$15.00
DOI: 10.1145/3106426.3106459

1 INTRODUCTION

Social media has evolved to become a source of varied kind of information [1]. The use of social media analysis to derive knowledge of its data to make smart decision is one of the most interesting and important research activities. Topic and opinion detection in social media facilitates the identification of emerging societal trends, analysis of public reactions to policies and business products. It provides a fast and reliable way of transforming a set of unlabelled documents into a well-structured knowledge base. Social media analysis covers a large set of online forums, Twitter, Facebook, blogs or other publicly available text streams are tracked and analysed. It is an increasingly popular platform for conveying opinions and thoughts, it seems natural to mine Twitter for potentially interesting trends regarding prominent topics in the news or popular culture [14].

Twitter is one of the most prominent renowned social media network. It is a global, public, distributed and real-time social and information network in which users post short messages called tweets [21]. A tweet is a short 140-character message. Registered users can read and post tweets, but unregistered users can only read those. In recent years, Twitter has been used as an ideal sources for spotting the information about societal interest and general peopler's opinions. Twitter has been seen as a potential new form of eWOM (electronic word-of-mouth) marketing by the businesses and organisations concerned with reputation management [9]. It also has been used as online surveillance platforms for assessing population-wide sentiment about public health issues like vaccines [4, 22] and tracking public health trends such as influenza outbreaks [16].

Although sentiment analysis often focuses on reviews of movies or consumer products, these probably form a tiny fraction of the social media. The remainder includes many friendly exchanges in social network sites, discussions of politics, sports, and the news in blogs and online forums as well as comments on media published in Tweeter, Facebook, YouTube, Flickr, Myspace, and Instagram [8]. Twitter, a micro-blogging social media service, is among the most pervasive social media services [2]. A successful sentiment classification model based on the expansive Twitter data could provide unprecedented utility for businesses, political groups and curious Internet users alike.

In this study, we focus on sentiment analysis and context-based topic mining on data analytics on Twitter data for for an election event in politic domain. A social media analytic model is proposed, which consists of two components; sentiment analysis and context-based topic modelling. The former component adopts a sentiment analysis model to evaluate public opinions; the latter components uses the Latent Dirichlet Allocation (LDA) model to study the related topics in public discussions on context basis. As problem ground, the case of 2010 Australian election for Prime Minister is studied using the proposed social media analysis model. We found the study results are interesting and the proposed model is promising. Our work reported in this paper delivers a two-fold contributions, as outlined below.

- As a theoretical contribution we proposed a high-level social media analysis model that helps improve our accessibility to public opinions and preferences;
- As a methodological contribution the proposed model will help reduce the dimensionality in social media for focused analytics.

The remainder of the paper is organised as follows. Section 2 provides a brief review of the related works. In Section 3, the proposed social media analytic model is introduced with technical details. The study on the election problem is discussed in Section 4 with related empirical and statistical analysis. Finally, concluding remarks are sketched in Section 5.

2 RELATED WORK

2.1 Opinion Mining

Opinion mining (a.k.a sentiment analysis) is the computational study of opinions, sentiments and emotions expressed in text [13]. The field of sentiment analysis and opinion mining is well-suited to various types of intelligence applications. Indeed, business intelligence seems to be one of the main factors behind corporate interest in the field [15]. Approaches for opinion mining can be broadly classified into machine learning-based method and lexicon-based method. Machine learning based approaches for opinion mining are supervised learning task. They utilise textual feature representation coupled with classification algorithms to infer the opinions expressed in the text [4, 22]. Unsupervised lexicon based techniques rely on the assumption that the collective polarity of a sentence is the sum of polarities of the individual words or phrases of that sentence [11]. Both [19] and [23] proposed unsupervised lexicon based systems. The authors of [19] used Word Net to classify the text using an assumption that with similar polarity have similar orientations but the authors of [23] have used Wilson opinion lexicon list's prior polarity lexicon subjectivity clue to quantify the semantic orientation of words by giving each type of word a numeric score. The authors of [17] have used manually created dictionary by hand tagging all adjectives found in their development corpus. All of them did reasonable pre-processing on their datasets. The Wilson lexicon based model outperformed WordNet based and manually created lexicon based system in speed even though the accuracy was similar.

Opinion mining in Tweeter is different from opinion mining from blogs and product review due to the size of the text. It is hard to design a system that analyses sentiment out of Twitter data quick, when Twitter users love informal use of English language, use of acronym, hashtag, innovatively-spelt word. This informal use of language is evolving every day, making it even harder. Unwanted data is hidden in big dataset. This unwanted data can impact the accuracy and efficiency of the analysis system. It can make the visualisation of data way bigger to present.

2.2 Topic Modelling

Topic modelling is a kind of text mining, a way of identifying patterns in a corpus or dataset. It is an attempt to inject semantic meaning into vocabulary [5]. After selecting a corpus and then run it through a tool which groups words across the corpus into topics. It is a method for finding and tracing clusters of words "topics"in short) in large bodies of texts and then group words across the corpus into topics. A topic modelling tool looks through a corpus for these clusters of words and groups them together by a process of similarity. In a good topic model, the words in topic make sense, for example "army", "tank", "captain" and "wheat", "farm", "crops".

Topic modelling itself is a powerful technique, when it is combined with opinion mining techniques it can be more useful by helping categorise large dataset and detecting underlying hidden pattern in data groups. Topic modelling tool like MALLET can be used to improve the process of opinion mining and sentiment analysis. It can be used to find hidden pattern in data that might reveal new knowledge, and can also be used to divide data into groups that might help discarding unwanted group of data.

Topic models are a useful and ubiquitous tool for understanding large corpora [7]. A topic model is a useful mechanism for identifying and characterising various concepts embedded in a document collection allowing the user to navigate the collection in a topic guided manner. According to [18] topics made up of significant words, provide the user with an overview of the content of the document collection. Each document is represented as a mixture of automatically constructed topics and the user may select documents related to a specific topic of interest and vice versa. Similarities between documents may be found by looking at what documents are assigned to a specific topic enabling the user to find other documents related to a given document. This methodology enables users to digest a larger number of documents, assisting them in spending more of their time in reading than finding relevant information.

Topic modelling technique and opinion mining technique has been used in the method proposed by [12] where they showed a novel probabilistic modelling framework based on LDA, called joint sentiment/topic model (JST), which detects sentiment and topic simultaneously from text. Unlike other machine learning approaches to sentiment classification which often require labelled corpora for classifier training, the proposed JST model is fully unsupervised. This work is somewhat close to our work, as they did use topic modelling to find hidden group of data in dataset, but their sentiment mining is on topic level not on entity level of tweets.

As what revealed from these discussions, quick and smart decision making ability is the key to success, and such key could be gained by adopting the techniques of sentiment analysis and context-based topic modelling. Our work presented in this paper was then motivated by the demand of such a key.

3 SOCIAL MEDIA ANALYTIC MODEL

3.1 High-level Architecture

The proposed social media analysis model comprises two parts; sentiment analysis model and context-based topic modelling. Twitter offers an easy way to access via Application Programming Interface (API), which can be used to interact with the service very easily. The set of the tweets is downloaded from Twitter via its API. The data set is first pre-processed, including the tasks such as data cleaning to prune away noisy data (e.g., punctuations and symbols), removing stopwords, and stemming words. The dataset is then analysed in two modules, sentiment analysis and context-based topic modelling. Figure 1 illustrates the framework of the proposed model.

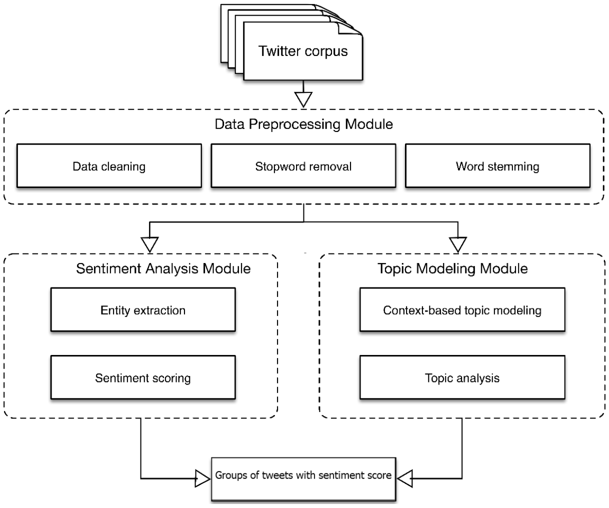

Figure 1: High-level Architecture

3.2 Sentiment Analysis

As shown in the Fig. 1, at the first stage, the Tweet Sentiment Analysis Model (TSAM) developed by [23] will be used to sentiment calculation and scoring for the entity. Due to space limit, we briefly described TSAM here and more detailed sentiment analysis process can be found in [23].

3.2.1 *Features Extraction.*

In TSAM model, instead of using all the words appearing in the news articles or tweets, the TSAM only extracted the opinion-bearing words as the features to input into opinion mining algorithm. Opinion words that are primarily used to express subjective opinions in the opinion sentence are identified and extracted. Words that encode a desirable state (e.g. beautiful, awesome) have a positive orientation, while words that represent undesirable states have negative orientations (e.g. disappointing, awful).

To identify the opinionated words, Wilson opinion lexicon list [20] was used to decide the words' semantic orientations. This list is a lexical resource of sentiment information for words, where each word is associated with positive, negative and neutral sentiment information. In this project, only the prior polarity lexicon subjectivity clue is used. The semantic orientation of words were quantified by giving each type of word a numeric score. Therefore, a positive and strong subjectivity words is assigned the semantic orientation score of +1, a positive and weak subjectivity word is assigned the semantic orientation score of +0.5, and a negative and strong subjectivity word is assigned the semantic orientation score of −1, a negative and weak subjectivity word is assigned the semantic orientation score of −0.5, and a neutral word is given the semantic orientation score of 0. These text strings can be placed into categories (positive, negative, neutral) and one can differentiate their strength or impact by assigning different weights. For example, the word "bankruptcy" can carry a stronger weight value than "lawsuit" even though they both might fall under the category "Negative".

Given a set of tweets, T, that contains a set of sentences, $T = \{s_1, s_2, ..., s_i\}$; and each sentence s_i describes something on a subset of entities $e = \{e_i...e_j | e_i, e_j \in E\}$, where E is the set of all entities. An entity can be a person, an organisation, a location, a product, an event, etc. Each sentence also contains a set of opinion word, w_k, $s = \{w_1, w_2, ..., w_l\}$. First, a Sentence Sentiment Scoring Function (SSSF) is used to determine the orientation of sentiment expressed on each entity e_i in s (i.e.,the pair of (e_i, s)). Then an Entity Sentiment Aggregation Function (ESAF) is used to obtain the total sentiment scores for a given entity e_i [23].

3.2.2 *Sentence Sentiment Scoring Function.*

In this stage, the classification algorithm detects all words that belong to Wilson lexicon list and extracts their polarity. Adjectives are good indicators of sentiment and have been used as features for sentiment classification by a number of researchers [10], [6]. However, it does not necessarily imply that other parts of speech do not contribute to expressions of opinion or sentiment. In fact, nouns (e.g., "gem") and verbs (e.g., "love") can be strong indicators for sentiment. Therefore, in this study, we use all the parts of speech. We summed up the semantic orientation score of the opinion words in the sentence to determine the orientation of the opinion sentence. The score function for a sentence is as follow:

$$score(s) = \sum_{(w_j : w_j \in s \wedge w_j \in WL)} \frac{w_j \cdot sentOri}{dis(w_j, e_i)} \tag{1}$$

where w_j is an opinion word, WL is the set of all opinion words from Wilson lexicon list and s is the sentence that contains the entity e_i, and $dis(w_j, e_i)$ is the distance between entity e_i and opinion word w_j in the sentence s, and $w_j.sentOri$ is the semantic orientation of the word w_j (i.e., +1, or +0.5, or 0, or −1, or −0.5). If a sentence contains more than one entity then the opinion word close to the entity has smaller value of $dis(w_j, e_i)$ and indicates this word makes more contribution to that entity's sentiment scores.

The *scores*(s) is normalised by the number of the opinion words, n, in the sentence to reflect the sentiment scores distributions of opinion words. So, normalised sentiment score will be:

$$score(s)_N = score(s) \div n \tag{2}$$

3.2.3 *Entity Sentiment Aggregation Function.*

In the given set of tweets, an entity appears in the set of sentences $s = \{s_1, s_2, ..., s_i\}$. We use co-occurrence of an entity and a sentiment word in the same sentence to mean that the sentiment is associated with that entity. This is not always accurate, particularly in complex sentences. Still the volume of text we process enables us to generate accurate sentiment scores.

For a given entity e_i, which may appear in multiple sentences $\{s_1, s_2, ..., s_i\}$, the normalised sentiment score for this entity in sentence s_k is $score(e_i, s_k)_N$. The total sentiment scores of this entity will be aggregated by Entity Sentiment Aggregation Function that is depicted as below:

$$score(e_i) = \sum_{(s_k : s_k \in s)} score(s_k)_N \tag{3}$$

This score is normalised by the number of the sentences, m, and then the final sentiment score for an entity will ranges in the interval $[+1, -1]$.

$$score(e_i)_N = score(e_i) \div m \tag{4}$$

In regard to sentiment intensity (or strength) for a given entity, e_i, appears in the sentences, the following heuristic rule is applied:

$$intensity(e_i) = \begin{cases} \text{SP} & \text{if } (+0.5 < score(e_i)_N < +1) \\ \text{P} & \text{if } (0 < score(e_i)_N < +0.5) \\ \text{Neu} & \text{if } (score(e_i)_N = 0) \\ \text{Neg} & \text{if } (-0.5 < score(e_i)_N < 0) \\ \text{SN} & \text{if } (-1 < score(e_i)_N < -0.5) \end{cases}$$

- *SN (Strong Negative)* Sentences about the entity e_i contain purely negative words or phrases or only allowed a slightly positive word.
- *N (Negative)* Sentences contain mainly negative phrases and words. There may be a few positive words, but the negative words or phrases outweigh the positive ones.
- *Neu (Neutral)* Sentences have a mediocre or balanced sentiment. The positive and negative words or phrases seem to balance each other, or it is neither positive nor negative overall. Even if there are more negative phrases, the positive ones use a stronger language than the negative ones.
- *P (Positive)* Sentences have mainly positive terms. There may be some negative ones; however, the positive ones are stronger and outweigh the negative ones.
- *SP (Strong Positive)* Sentences have purely positive words expressing strong affirmative feelings with no complaints. It may have the smallest negative words, but the sentence has mostly great-sounding words or phrases.

3.3 Context-based Topic Modelling

At the second stage, the results from the TSAM will be used by Topic Modelling component for the categorization process of tweets for hidden community detection. In this research project, the Latent Dirichlet allocation (LDA) model, a state-of-the-art technique for context-based topic modelling is adopted [3]. In practice, we utilized MALLET, one of the most popular and freely available text mining tools to handle the task of context-based topic modeling. MALLET is an open source Java-based package for statistical natural language processing applications to text. It includes an extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyper-parameter optimization, and tools for inferring topics for new documents given trained models. In our tweets dataset, there are many types of entity ("position", "company", "place" etc.) beside person (Tony Abbott, Julia Gillard, Rick James, Kevin Rudd etc.). Our target entities are two Prime Minister candidates (Julia Gillard and Tony Abbott) only. Therefore, by using the LDA model, three topic models were produced and they are two topic models of two prime minister candidates and one topic model that is not related to the prime minster candidates. The words in these topic models are used to categorize tweets to find underlying hidden communities and patterns.

For example, Table 1 shows three topic models. Each topic model contains a set of words shown in column one, two, and three, respectively. The number of word a tweet has in common in a topic model is the score of that tweet against the three topic models. If a tweet has five common words in Topic Model 0, two common words against Topic Model 1, and one common word against Topic Model 2, then that tweet falls under Group 0. The highest number of common words against a topic model decides which group the tweets belong to. If there are two even scores against two topic models then that tweet belongs to both groups of tweets.

Table 2 shows score against the topic models. The line ID (Line#) and the tweet sentence with entity are shown in column one. The scores for Topic Model 0, Topic Model 1 and Topic Model 2 are diplayed in column two, three, and four, respectively.

The community of tweets named Group 0 holds most of the tweets about Tony Abbott, the community named Group 2 holds most of the tweets about Julia Gillard, and the community of tweets named Group 1 holds most of the tweets about the others. As we are more interested in two prime minister candidates Tony Abbott and Julia Gillard, we can discard the community of tweets named Group 1. This discovery of detecting hidden community of data in our dataset inspired us to do further topic modelling on the categorized groups to go deeper to see any more underlying hidden community can be discovered. Interestingly, we found one useful pattern in Group 2, the pattern gives us more insight into the data to sub-categorize this group and make deeper meaningful sense that could not have been seen before. The sentiment scores obtained from the first stage has been injected to the communities of tweets to see public opinion polarity on the entities in the community.

In search of detecting underlying hidden pattern or community of tweets in these newly categorized three groups of tweets, the further topic modelling on each group of tweets were performed. Hence, two sub-topic models from each group of tweets were created.

After categorising two topic models of Group 02, we found hidden communities of tweets; one community of tweets predominantly talks about one thing while the other community of tweets talk about something else. This is an important finding; opinion mining technique alone is not capable of achieving this; it needs to couple with topic modelling technique to do it.

Table 1: Topic Models with set of the words included in each model produced by LDA

Topic Model 0	Topic Model 1	Topic Model 2
0.00432	0.00317	0.00405
ausvotes abbott person	ausvotes female	ausvotes gillard julia
company abc election tony	technology person	minister house position
streaming cassidy crabb	censorship internet	person prime government
coverage live twitter	mandatory position bit	election car call way
technology green uhlmann	james history rick going	arrives industryterm
news bit online media	election punter happened	abc news driving govt

Table 2: Tweet's matching word Score against the Topic Models

Line# (Tweet#)	Topic Model 0	Topic Model 1	Topic Model 2
Line# 1 Tony Abbott: I regret to say that we broke the faith with the Howard battlers #ausvotes|Tony Abbott;Person	5	2	2
Line# 2 @theburgerman: Work Choices not only dead but cremated says Tony Abbott #ausvotes|Tony Abbott;Person	6	2	2
Line# 3 @skynewsaust: Watch Tony Abbott online here at a Qld LNP conference #ausvotes|http;Technology	6	2	2
Line#29 so my first federal election and I m going to be 15000km around the world don t like! #ausvotes|federal election;PoliticalEvent	8	2	2
Line# 46 COME ON Switching off Rick James for this? This is why democracy SUCKS #ausvotes|Rick James;Person	3	4	3
Line# 57 Graham Richardson says Gillard s Aussie enough to walk into a bar anywhere and "charm em" #ausvotes|Graham Richardson;Person	2	6	2

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Data Source and Preprocessing

The original dataset used in this work is comprised of 57000 Australian tweets of two weeks period during Australian Federal Election 2010. When the election date 21 August 2010 was announced on 17 July 2010, Twitter has seen a surge in tweets. This dataset is specifically comprised of 2 weeks of tweets from Saturday 17 to Saturday 31 July. The tweets with hashtag *#ausvotes* have been considered only. Even though there could be tweets about the election without that hashtag. We are missing out the sentiment score of those tweets. The collected tweets data were split into 57 files each containing about 1000 tweets that make a total of 57000 tweets. All the tweets were chronologically ordered by tweet id within the dataset for the sake of the experiment.

One non-trivial task of tweets data collection for sentiment analysis is the extraction of the relevant entities from the tweets. To identify and extract the entities that appear in the sentences the Open Calais was used currently. Open Calais is one of well-known entity extraction tools. Open Calais extracts entities from textual (natural language) input and returns an XML document contains meta-information about entities in RDF format, including name and type. Detailed information can be found at http://www.opencalais.com.

The original 57000 tweets were used in the first stage. After the first stage, there are 457 remained relevant tweets with their sentiment scores and entity tagged for the second stage. The second stage has implemented topic modelling technique on the remained dataset to discover the underlying hidden pattern to better understand what public is trying to say.

The data preprocessing techniques including stop word removal, punctuation and symbol removal, stemming were applied to the tweets dataset. The Table 3 shows which parts were removel from the tweets.

We measure the system performance with its accuracy as following:

$$accuracy = \frac{\texttt{NTSCL}}{\texttt{TNTTS}} \tag{5}$$

where NTSCL is the number of tweets the system correctly labelled and TNTTS is the total number of tweets in a test set.

4.2 Results of Categorised Dataset

Topic model can help categorise large dataset to discover unwanted group of data. In this study, tweet dataset is categorised into three topic models - Group 0, Group 1 and Group 2. Fig. 2, 3, 4 represent all the entities in Group 0, Group 1 and Group 2, respectively. It was found that most of the tweets in group 0 are about Tony Abbott, most of tweets in group 1 are about everyone other than Julia or Tony, and most of the tweets in group 2 are about Julia Gillard. In group 0, the target entity is the prime minister candidate Tony Abbott who got 56% positive sentiment, 38% negative sentiment, and 6% was neutral (see Fig. 2). In group 2, the target entity is the prime minister candidate Julia Gillard who got 60% positive sentiment, 24% negative sentiment, and 16% was neutral (see Fig. 4). In order

Table 3: Data preprocessing that removed unwanted parts from tweets

Sample	Type	Task
A, an, on, at, in etc.	Stopwords	These words do not bear any sentiment
RT	Retweet	Reposting another user's tweet
@	Mention	Tag used to mention another user
#	Hashtag	Hashtags are used to tag a tweet to a certain topic.
URL	URL	Typically a link to an external resource
, . ? ;	Punctuation	These do not bear any sentiment
&, *, \,), (, $	Symbol	These do not bear any sentiment

to do further analysis on Group 0 and Group 2 more efficiently, the Group 1 was discarded since it was an unwanted group.

In comparison of Fig. 2 and 4 it can be seen that Julia Gillard's 64% positive, 24% negative and 16% neutral is better than Tony Abbott's 54% positive, 38% negative and 6% neutral in terms of public sentiment polarity. This comparison suggests that Julia Gillard got more chance of winning the election that Tony Abbott and in reality this was the fact in the Australian Federal Election 2010.

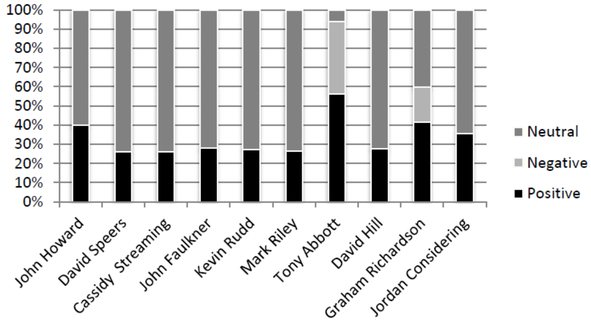

Figure 2: Group 0 Entity Sentiment Comparison in Percentage

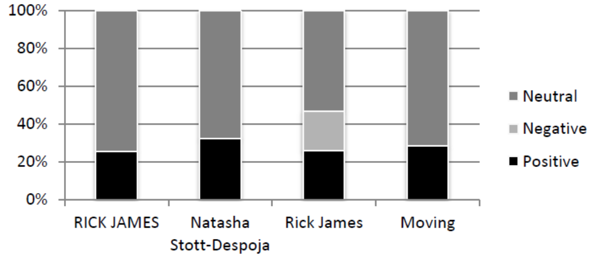

Figure 3: Group 1 Entity Sentiment Comparison in Percentage

Fig 5 is the overall comparison of all the entities that we had before doing topic modelling; the less interesting entities making the chart bigger. Fig 6 is the representation of the comparison of all the entities after doing topic modelling and discarding the less interesting group of entities; making the chart smaller to present without impacting the result. Even on this small dataset, we could remove 21% entity and after doing topic model and present a less crowded chart without impacting the result. Figure 5 is with all the entity, Figure 6 is a clearer version with only interesting topics and communities.

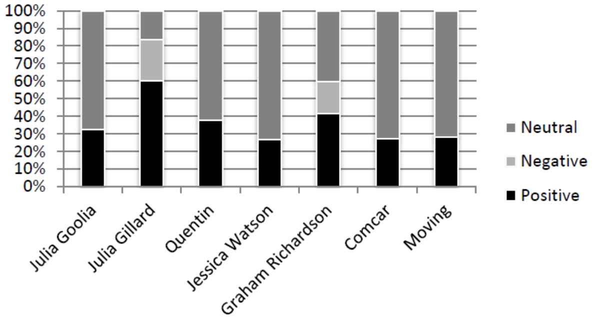

Figure 4: Group 2 Entity Sentiment Comparison in Percentage

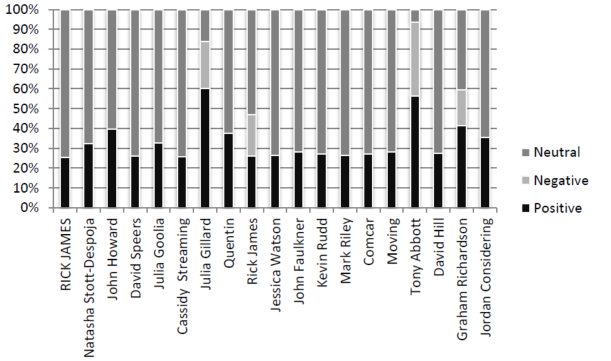

Figure 5: Chart with Unwanted Entity Group before Topic Modelling

4.3 Topic Modelling for Community Detection

The further topic modelling on the categorized groups was carried out in order to go deeper to discover any more underlying hidden community. Interestingly, we found one useful pattern in Group 2, where it gives us more insight into the data to sub-categorize this

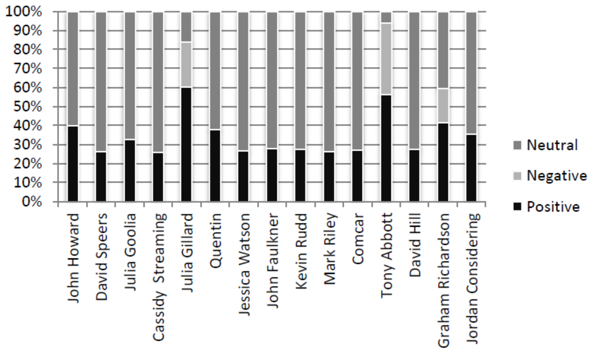

Figure 6: Chart after discarding Uninteresting Entity Group through Topic Modelling

group and make deeper meaningful sense that could not have been seen before. Fig. 7 is the result of doing further topic modelling on Group 2 (Julia Gillard group). The effort of doing more topic modelling on that group revealed a hidden community of tweets in that group. Sixty three percent of the tweets in that group were about Julia going to the Governor General (GG) and the other thirty seven percent were about other election interests.

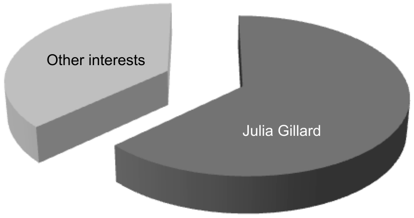

Figure 7: Hidden Sub-group in Julia Gillard's Group

4.4 Discussions

With the help of topic modelling, the unwanted topics, entities or tweets can be discarded. Hence the accuracy of calculation can be improved and the system can run faster as it will be handling less data when dealing with very large datasets. Topic model can uncover hidden group or pattern by grouping topics. For example, the dataset used in this study contained all tweets that have hashtag "#ausvote" so that the tweets are related to the election. After first attempt of topic modelling, we were able to divide the tweets in three groups. We did more topic modelling on the grouped data, and found pattern or hidden underlying community of tweets. This technique can be used to help federal election campaign. For example, in federal election campaign of 2016 LNP has mandates that will cause education and medicare cost go higher. In all the tweets with "#ausvote", if the topic modelling is used, it can help categorize, how many percentage of tweets of LNP are about those two unpopular mandates. This analysis can help LNP to change their mandate or advertise the reason why they are doing it to educate people to make them understand that it is necessary for the economy of the country and to win more votes to win the election. On this project's dataset (Australian Federal Election 2010), after doing topic modelling, it is visible that a big group of tweets about Julia Gillard was to do with her going to the Governor General. This shows that it is possible to uncover hidden group and interpret meaning of that group.

This technique of uncovering hidden group can also be used to uncover imminent criminal or terrorists' activity that the criminals are planning by topic modelling the communication data on the internet. This use of topic modelling can be hugely beneficial to the society as it will save lives from death and injury and property from destruction. Also this kind of work will not require much to invest, so it is a very economical way to fight the war on terrorism or big scale criminal acts.

By uncovering hidden group and categorising data from very large datasets, topic model can improve the speed and accuracy of the result. Our dataset is not a very large dataset. The benefit of topic modelling increases with the size of the dataset. If we analyse our result, we would still see the aforementioned benefits in little scale. Our target is to compare the two Prime Minister candidates Julia Gillard and Tony Abbott. After calculating the sentiment and tagging the entity of the tweets, we did topic modelling and found a hidden group that can be discarded without impacting our result. If it were a huge dataset with thousands of entity, it would make really cumbersome to represent them in one chart. By detecting the less interesting entity group, we can remove them from the chart and produce a smaller and easily presentable chart without impacting the result. If we did topic modelling before sentiment calculation, we definitely could get away with less calculation by removing less interesting groups. Less calculation means less use of computing processing that could lead to quicker output and less energy cost to save direct financial cost, making the system more economical and efficient.

5 CONCLUSIONS

This study has demonstrated that topic modelling technique can be utilised on opinion mining process to enhance the feature of the system. In contrast to most of the existing research works in this field of opinion mining, which tend to overlook detecting the underlying hidden group of data in the dataset, this work focused on detecting the hidden pattern in dataset. The experimental result shows topic model can uncover hidden community of tweets by categorising tweets in groups, and further topic modelling on a group of tweets revealed an underlying pattern deeply hidden in that group of tweets. This topic modelling technique made TSAM system even better and this technique certainly can be used to enhance the performance of any opinion mining or sentiment analysis systems; whether it is lexicon based unsupervised, supervised system or corpus based supervised system or a hybrid system. The outcomes of this work are the ability to categorise data in a meaningful way and to find underlying hidden group of data. These outcomes will enable us to make informed decision and make prediction on emerging social events.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 30–38.

[2] Amir Asiaee T, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1602–1606.

[3] Michael I. Jordan; David M. Blei, Andrew Y. Ng. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(1) (2003), 993–1022.

[4] Adam G Dunn, Julie Leask, Xujuan Zhou, Kenneth D Mandl, and Enrico Coiera. 2015. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *Journal of medical Internet research* 17, 6 (2015).

[5] Shawn Graham, Scott Weingart, and Ian Milligan. 2012. Getting started with Topic Modeling and MALLET. *The Programming Historian* 2 (2012), 12.

[6] Vasileios Hatzivassiloglou and Janyce M Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 299–305.

[7] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. ACM, 80–88.

[8] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 607–618.

[9] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci.* 60, 11 (2009), 2169–2188.

[10] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 1115–1118.

[11] Chetan Kaushik and Atul Mishra. 2014. A scalable, lexicon based technique for sentiment analysis. *arXiv preprint arXiv:1410.2265* (2014).

[12] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 375–384.

[13] Bing Liu. 2010. *Handbook of Natural Language Processing* (second edition ed.). Chapter Sentiment Analysis and Subjectivity.

[14] Ravi Parikh and Matin Movassate. 2009. Sentiment analysis of user-generated twitter updates using various classification techniques. *CS224N Final Report* (2009), 1–18.

[15] S. Shahheidari, H. Dong, and M. N. R. B. Daud. 2013. Twitter Sentiment Mining: A Multi Domain Analysis. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on*. 144–149. DOI: http://dx.doi.org/10.1109/CISIS.2013.31

[16] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one* 6, 5 (2011), e19467.

[17] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[18] JW Uys, ND Du Preez, and EW Uys. 2008. Leveraging unstructured information using topic modelling. In *Management of Engineering & Technology, 2008. PICMET 2008. Portland International Conference on*. IEEE, 955–961.

[19] Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for Sentiment Analysis in Twitter.. In *COLING*. 2345–2354.

[20] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 347–354.

[21] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. 2014. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1907–1916.

[22] Xujuan Zhou, Enrico W Coiera, Guy Tsafnat, Diana Arachi, Mei-Sing Ong, Adam G Dunn, and others. 2015. Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter.. In *MedInfo*. 761–765.

[23] Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. IEEE, 557–562.