World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:7, 2018

Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis

Sidi Yang, Haiyi Zhang

***Abstract*—**Twitter is a microblogging platform, where millions of users daily share their attitudes, views, and opinions. Using a probabilistic Latent Dirichlet Allocation (LDA) topic model to discern the most popular topics in the Twitter data is an effective way to analyze a large set of tweets to find a set of topics in a computationally efficient manner. Sentiment analysis provides an effective method to show the emotions and sentiments found in each tweet and an efficient way to summarize the results in a manner that is clearly understood. The primary goal of this paper is to explore text mining, extract and analyze useful information from unstructured text using two approaches: LDA topic modelling and sentiment analysis by examining Twitter plain text data in English. These two methods allow people to dig data more effectively and efficiently. LDA topic model and sentiment analysis can also be applied to provide insight views in business and scientific fields.

***Keywords*—**Text mining, Twitter, topic model, sentiment analysis.

Open Science Index, Computer and Information Engineering Vol:12, No:7, 2018 waset.org/Publication/10009246

I. Introduction

SOCIAL networks such as Twitter, Facebook, and LinkedIn have become crucial sources of social information and have attracted interest and curiosity in social research and commerce. In addition, methods of acquiring and analyzing the huge amounts of data generated by social media have in themselves become the interest of researchers. As a social online microblogging service, Twitter allows users to broadcast and interact with posts, called "tweets". Twitter has rapidly obtained worldwide popularity since its launch in 2006. According to a report by *Insights West 2017 Canadian Social Media Monitor*, 29% of Canadian adults use Twitter, and over half of them are daily users [1]. This fact makes Twitter as a rich source for capturing information. In this paper, the efficacy and effectiveness of two text mining techniques are applied to analyse Twitter plain text data.

II. Background

A. Text Mining

Text Mining refers to the process of extracting high-quality information from a large amount of unstructured text using computational methods and techniques. Unstructured data are ubiquitous and can be in forms such as new articles, books, and social media. The amount of unstructured text data is growing rapidly, and Computer World magazine declares that unstructured information might be more than 70%-80% of all data in organisations [2]. Hence, text mining is relevant to enable the effective and efficient use of huge quantities of text.

Sidi Yang and Haiyi Zhang is with the Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada (e-mail: 123278y@acadiau.ca, haiyi.zhang@acadiau.ca).

B. Topic Model

Topic modelling is an essential algorithm used in text mining. A topic model is a probabilistic model that discovers the main themes in a collection of documents. The basic idea is to treat the documents as mixtures of topics in the topic model, and each topic is viewed as a probability distribution of the words. When using a topic model as a text-mining tool, each topic is viewed as a collection of words, and each document can be viewed as a set of topics with different proportions depending on the frequency that terms appear.

The topics are what the documents talk about. Intuitively, a document is about a topic, and the topic words tend to appear more often than other words [3]. For example, considering the themes in a collection of recipes in a cookbook, words like "sugar", "teaspoon", "oil" will appear more frequently, while in technical IT documents, words like "computer", "research", "algorithms" will appear more often. In addition, words like "is", "are", "a" will appear in all the documents, regardless of the themes, and the method of removing these kinds of words are called “stopwords”. “Stopword” is discussed in the section on data preprocessing.

1. Bag-of-Words Assumption

A Bag-of-words model represents the “attendance” of the words in a text document. Regardless of the structure of the terms, or the grammar of a sentence, the text document is thought of as a "bag" where syntax is not considered. The only thing that matters in the bag-of-words model is whether a word appears in the text, not where and how.

2. Latent Dirichlet Allocation (LDA)

LDA is a commonly used technique in topic modelling. David Blei, Andrew Ng and Michael I. Jordan first proposed LDA as a topic discovery graphical model in 2003 [4]. The basic idea behind LDA is that documents exhibit multiple topics. The topics, using the bag-of-words assumption, are formally defined as a distribution over a fixed vocabulary [5].

To summarise how the topics are inferred and to achieve an LDA model from the documents generated, the generative process is described as [6]:

1. For each topic: Decide what words are likely.
2. For each document:
   a. Decide the proportions of topics that should be in the document,
   b. For each word:
      i. Choose a topic,

Open Science Index, Computer and Information Engineering Vol:12, No:7, 2018 waset.org/Publication/10009246

ii. Given this topic, choose a likely word (Generated in step 1).

C. Sentiment Analysis

Opinions influence human behaviours such that our thoughts or choices can depend on the evaluation by other people. The study of sentiment analysis contains the subjects, includes opinions on these subjects, and the related conceptions of these subjects. Sentiment analysis has grown into one of the most active areas in natural-language processing (NLP) and data mining.

Sentiment analysis is broken into five steps. A typical model is shown in Fig. 1.

Data Collection is collecting user-generated data, which can involve data from social media networks and blogs, etc. This data is generally unstructured, expressed in various ways by word choices, idiomatic phrases, and so on.

Text Preparation is the preparation step before data extraction. Irrelevant contents and non-textual data are eliminated.

Sentiment Detection examines the opinions in the extracted text. The contents with opinions are retained, and the contents with factual information are discarded.

Sentiment Classification is to apply some approaches to classify the sentiments. Typical approaches include the machine learning approach, the sentiment orientation approach, and a lexicon-based approach.

Presentation of Output is to display the results in a direct way such as in graphs (pie charts, line graphs, and so on).

III. Experiment Design and Implementation Approaches

A. Overview

Thor: Ragnarok is an American hero movie released in Canada and United States on November 3, 2017. Our experiment is to analyse audience reviews of this movie on the released day. Tweets from 10 p.m. (UTC -0400) November 3 to 10 p.m. (UTC -0400) November 4 are fetched. The total number of retrieved tweets was 185,185. The basic workflow chart of this experiment is shown in Fig. 2.

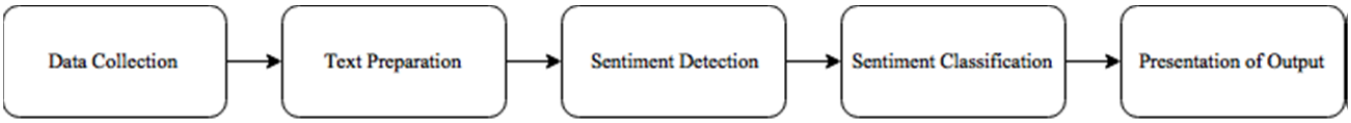

Fig. 1 Sentiment Analysis Model

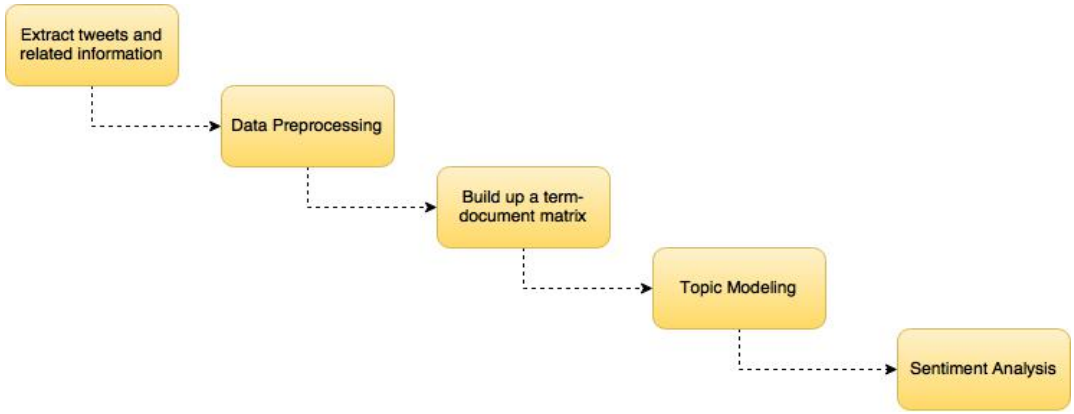

Fig. 2 The basic workflow of the experiments

B. Coding Language and Environment

R is an open source programming language, and it runs on a variety of platforms including Windows, MacOS, and UNIX. The R programming language is a simple, well-developed effective programming language which has the functionality of loops, conditionals, user-defined recursive functions, and input and output facilities [7]. Most classic statistics problems and many of the latest methodologies are available in R. Many standard statistical techniques are built-in to R, but many others are provided as packages. R packages are the fundamental units of user-shared codes and are developed in R and sometimes in C, C++, Java, and Fortran. Some R packages are used in the experiment, including the **twitteR** package (collecting tweets and corresponding information from the Twitter website and offers access to the Twitter web API), **tm** package (preprocessing, clustering, categorization, summarization, and association), **topicmodels** package (fitting models based on the data structures from the tm package) and **syuzhet** package (sentiment analysis).

C. Data Retrieving and Preprocessing

1. Data Retrieving

In general, a tweet by any user is public and readable by anyone by default. Registering an account on the Twitter developer platform provides access to all the tweeted data. ***FilterStream*** is used to retrieve tweet data in the experiment. This function is provided *streamR*, which connects to the Twitter Streaming API and opens a stream to capture the results based on specific conditions, such as keywords and locations.

Open Science Index, Computer and Information Engineering Vol:12, No:7, 2018 waset.org/Publication/10009246

2. Data Preprocessing

Most of the Twitter data are highly unstructured. There could be typos, slang usage, and grammar mistakes. Then cleansing steps are applied to the documents to generate structured data.

- **Convert to lower case letters.** Conversion into lower case letters is necessary because text analysis is case sensitive. In the probabilistic model, the frequency of each letter is counted so that "Text" and "text" are treated as different words if the case conversion is not applied. All the letters are converted into lower cases.
- **Remove @user and links.** Twitter has special rules regarding reserved symbols and links that could cause confusion in later analysis. The @ symbol is used when the user mentions some other user to read this tweet.
- **Remove punctuations and digits.** This is a general step used in many text mining techniques. Punctuation in sentences makes the text more readable for humans, but a machine does not distinguish punctuation and digits from other characters. Punctuation is removed because text analysis is not concerned with the digits. Numeric digits usually do not influence the meaning of the text.
- **Remove stopwords.** Stopwords refer to the words which usually have no analytic value, words such as 'a', 'and', 'the' etc. These words make the sentences more readable to humans, but confound the analysis. Words can be added to the list of stopwords depending on the specific requirements.
- **Remove extra white spaces.** The earlier preprocessing steps can generate extra white spaces. Removing the extra spaces is necessary in text cleansing.
- **Stem Documents.** Some words in the text have the same meaning but in different forms. Stemming is the process of eliminating affixes from words to convert the words into their base form; for example, stemming "run", "runs" and "running" into "run".

3. Sentiment Analysis Preparation

The ***syuzhet*** package is used for sentiment analysis. This package extracts sentiment and sentiment-derived plot arcs from the text [8]. The dictionaries incorporate four optional sentiment lexicons, and a lexicon named "***nrc***" lexicon is used in this paper for sentiment analysis. The "***nrc***" lexicon was developed by Saif Mohammad and Peter Turney [9]. The NRC Emotion Lexicon involves a list of English words associated with eight emotions – anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. In addition, there are two sentiments: negative and positive.

D. Top Frequent Words

After text cleansing, the corpus that contains 185,185 collected tweets. The next step is to search for the most frequent words in this corpus to find out what people are most interested in about this movie. Table I shows the top 20 frequent words in these tweets.

TABLE I
THE TOP 20 FREQUENT WORDS IN THE CORPUS

word	thor	thorragnarok	ragnarok	movie	watch	marvel	loki	get	hela	good
freq	107611	60427	54253	17426	12408	10754	10538	9004	6622	6588
word	thank	weekend	like	film	best	cast	hulk	love	ticket	review
freq	6566	6532	6395	6332	6264	5890	5149	5037	4756	4712

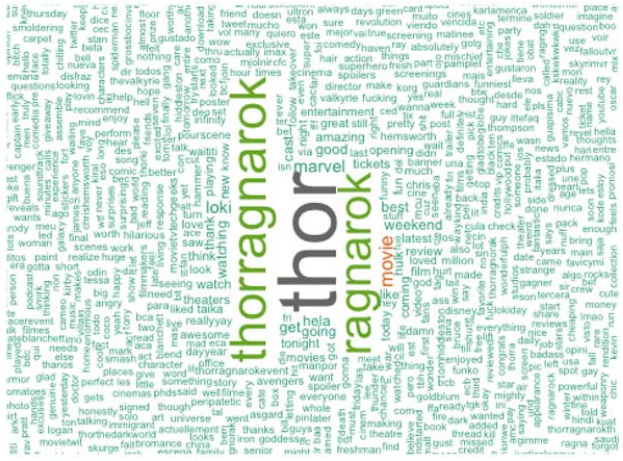

Fig. 3 Word cloud with minimum frequency = 50

9

The words "thor", "loki", "hela", and "hulk" are character names in this movie. The high-frequency words "good", "best", and "love" refer to a positive attitude. This result can be verified by using sentiment analysis. Fig. 3 shows all the words with a frequency higher than 50.

E. LDA Topic Modelling

This basic analysis gives us a brief understanding of the collected data. LDA topic modelling is applied to analyze the topics hidden in these tweets. K = 15 is used as the number of topics. The top 10 frequent words in each topic are in Table II.

The results lead to some observations about people's reactions to the film. In Topic 5, Topic 12, and Topic 14, the words like "jeff", "goldblum", "chrishemsworth", "tomhiddleston", "cateblanchett" show that people have discussed the cast (Jeff Goldblum, Chris Hemsworth, Tom Hiddleston, Cate Blanchett). The words "taika" and "waitit" in Topic 5 refer to the director of this movie (Taika Waititi). Topic 8, Topic 10, and Topic 13 show that people are interested in the movie, Avengers: Infinity War which is a sequel to Thor: Ragnarok to be released in 2018. In Topic 6, the Twitter users were concerned about the comparison between Thor: Ragnarok and a similar superhero movie entitled Justice League which was released around the same time as Thor: Ragnarok. Some other topics, for example, Topic 3, Topic4, Topic 9 and Topic 15 indicate people's review of Thor: Ragnarok.

Open Science Index, Computer and Information Engineering Vol:12, No:7, 2018 waset.org/Publication/10009246

TABLE II
TOP 10 WORDS IN EACH TOPIC

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
watch	thorragnarok	loki	good	thorragnarok
hela	cast	thor	really	movie
ticket	watch	disney	say	taika
play	night	read	great	everyone
theater	open	worth	pretty	waititi
tonight	fan	marvel	story	life
follow	surprise	film	comedy	goldblum
thorragnarok	screen	wonderful	thought	cinema
win	last	care	character	jeff
poster	filmmaker	age	actual	kirby
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
ragnarok	thor	like	weekend	thor
think	ragnarok	thor	thor	ragnarok
latest	hulk	ragnarok	watch	trailer
entertainment	join	video	wait	final
movietvtechgeeks	pick	spoiler	tomorrow	infinite
comic	revolution	scene	valkyrie	top
book	art	credit	exciting	hammer
thor	revenge	end	talk	war
justice	banner	fire	big	universe
league	beautiful	miss	ready	song
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
thor	thor	thor	thor	love
thank	ragnarok	show	need	fun
live	tomhiddleston	peripatetic	chris	amazing
every	hammer	word	hemsworth	great
definite	happy	bet	avenger	movie
review	lokiday	rock	cate	funny
perform	hair	spiderman	world	action
theatre	chrishemsworth	black	guardian	recommend
time	death	power	blanchett	laugh
imax	cateblanchett	avenger	funny	awesome

F. Sentiment Analysis

Section III.E showed that the total sentiment of the data could be positive because some high-frequency words in this corpus are positive. This idea can be demonstrated with the sentiment analysis of the movie tweet data in this section. The results of eight emotions and two sentiments are shown in Tables III and IV.

TABLE III
THOR: RAGNAROK DATA SENTIMENT ANALYSIS – EIGHT EMOTIONS

Emotion	Count
anger	119768
anticipation	73386
disgust	9072
fear	22506
joy	82982
sadness	135762
surprise	44116
trust	87389

TABLE IV
THOR: RAGNAROK DATA SENTIMENT ANALYSIS RESULTS – TWO SENTIMENTS (LEFT)

Sentiment	Count
positive	247017
negative	154653

The percentage of the total number of the positive terms in the collected data is 61.5%, which refers to an overall positive attitude towards this movie.

The results in this section can be compared to the reviews found on well-known movie review aggregation websites such as the Internet Movie Database (IMDb) and Rotten Tomatoes. Thor: Ragnarok was rated 8.1/10 by 185,663 users on IMDb [10]. On the Rotten Tomatoes website, it obtained 88% positive reviews by 84,275 users [11]. The results show that users from the IMDb and Rotten Tomatoes had more than 80% positive reviews for this movie. Limitations in my experimental data prevent me from making an exact comparison between the movie review website results and my results. For example, my experimental data were only collected for 24 hours, and the experimented data is only natural language text. Nevertheless, the website data qualitatively agrees with my results that the overall sentiment for this movie was positive.

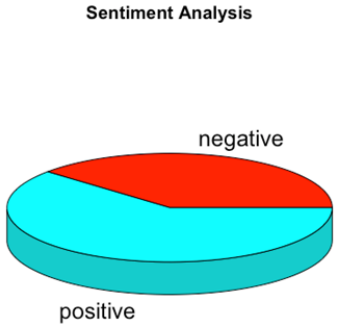

Fig. 4 Thor: Ragnarok data sentiment analysis pie chart – Two Sentiments (right)

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, the principles, related research field theories and the application of LDA topic modelling and sentiment analysis on Twitter text were discussed. Using an LDA topic model as a probabilistic model to explore hidden topics in documents makes it easy to analyze a large set of tweets and conclude what the main topics of the tweets were. Sentiment analysis provides a good method of showing the emotions and sentiments found in each tweet and of summarizing the results. Both LDA topic modelling and sentiment analysis are powerful exploratory tools when used on a large collection of Twitter data. However, there are still limitations in my experiments:

- Only plain text of Twitter is used for the text analysis. Videos links, emojis, and website links are removed during the process.
- Only English experimental data is examined.
- Because of limited computing power, the number of tweets is limited in my experiments. A very large dataset has the possibility of higher memory usage which could lead to the R session being aborted. Even with the size of the data sets in my experiments, computing time was time-consuming. For example, in the last experiment, retrieving the objective of 185,185 tweets took up to 24 hours and preparing the tweets for analysis took seven hours.

B. Future Work

Here are some thoughts for future work:

- The experiments in this thesis used plain text in English. Emoji symbols are removed during the data pre-processing step since they are ideograms not language. In fact, Twitter users use emoji very often in their posts. Emoji can be a good source for sentiment analysis. For example, the smiling face "😄" can represent joy and positive sentiment. In fact, each emoji has a unique code represented by Unicode format. In the R programming language, "😄"can be transferred into the plain text format "\U0001f604". Thus, exploring emojis and using them as data in sentiment analysis could be a valuable.
- LDA topic modelling and sentiment analysis are used as methods of text analysis. In the future, it would be interesting to test other methods of text mining such as LSA and PLSA.

REFERENCES

[1] 2017 Social Media Monitor | Insights West. (2017). Insightswest.com. Retrieved from https://insightswest.com/reports/2017-social-media-monitor.

[2] Chakraborty, Goutam. "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining" (PDF). SAS. Retrieved June 24, 2016.

[3] David M. Blei. *Probabilistic topic models.* In Communications of the ACM, volume 55, pages 77-84, 2012.

[4] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. *"Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5):* pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

[5] David M. Blei and John D. Lafferty. Topic models. In Text Mining: Classification, Clustering, and Applications, pages 71-94, 2009.

[6] Martin Ponweiser. Latent dirichlet allocation in r. Diploma thesis, Vienna University of Business and Economics, 2012.

[7] Venables, W. N., Smith, D. M., and R Development Core Team. (2010) An Introduction to R. Version 2.11.1, cran.r-project.org/doc/manuals/R-intro.pdf.

[8] Jockers, M. L. (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text. https://github.com/mjockers/syuzhet.

[9] Mohammad, S. and Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence, 29 (3), 436-465.

[10] Thor: Ragnarok (2017). (n.d.). Retrieved January 14, 2018, from http://www.imdb.com/title/tt3501632/?ref_=nv_sr_1.

[11] Pearson, E. (2018, January 08). Thor: Ragnarok. Retrieved January 14, 2018, from https://www.rottentomatoes.com/m/thor_ragnarok_2017.

Open Science Index, Computer and Information Engineering Vol:12, No:7, 2018 waset.org/Publication/10009246