Sentiment Clustering with Topic and Temporal Information from Large Email Dataset

Sisi Liu
Information Technology
James Cook University
Cairns, QLD 4870, Australia

Guochen Cai
Information Technology
James Cook University
Cairns, QLD 4870, Australia

Ickjai Lee
Information Technology
James Cook University
Cairns, QLD 4870, Australia

{Sisi.Liu1, Guochen.Cai}@my.jcu.edu.au; Ickjai.Lee@jcu.edu.au

Abstract

Sentiment analysis with features addition to opinion words has been an appealing area in recent studies. Some research has been conducted for finding relationship between sentiments, topics and temporal sentiment analysis. Nevertheless, Email sentiment analysis received relatively less attention due to the complexity of its structure and indirectness of its language. This paper introduces a systematic framework for sentiment clustering using topic and temporal features for large Email datasets. Interesting Email and sentiment distribution patterns are summarized and discussed with empirical results.

1 Introduction

The generation of enormous diversified data stream by social networking and communication contributes to the rapid development of text mining and its related area (Hao et al., 2013). Literature indicates that product reviews, Twitter corpus and news articles are common sources for conducting sentiment analysis (Ravi and Ravi, 2015), whereas Electronic mail (Email), as one of the most adapted means of communication and networking, is a rare option due to its complex structure and natural language characteristics (Tang et al., 2014). However, the efficiency, compatibility and ease of communication embed great business potential in Email messages (Tang et al., 2014), which is a promising and meaningful sentiment analysis subject.

Sentiment analysis is one of the most appealing areas in text mining among researchers. In the past few decades, sentiment analysis techniques, both machine learning approaches and statistical approaches, have improved significantly and been applied to various industries, such as stock market prediction, customer relationship management, and e-learning (Feldman, 2013; Liu, 2015; Ortigosa et al., 2014; Smailović et al., 2013). Herein, some researchers extend their studies to enriching sentiment analysis by adding additional features. For instance, Mei et al. (2007) propose a novel topic-sentiment mixture model using probabilistic testing for topic and sentiment discovery; Saif et al. (2012) show that adding semantic features results in more accurate sentiment classification. Additionally, Fukuhara et al. (2007) introduce the idea of generating time and sentiment graph using Dice coefficient probabilistic model. However, no qualitative and quantitative experiments have been undertaken for the evaluation of the proposed method.

This research paper develops a systematic scheme of approach for discovering sentiment distribution patterns from large Email corpus based on clustering results of topic and temporal information using bag-of-words model as distance matrix and DBSCAN (Ester et al., 1996) algorithm for clustering and pattern analysis, addressing the following contributions:

a) introducing a systematic scheme of approach composed of bag-of-words term weighting method and DBSCAN clustering algorithm for Email sentiment pattern discovery using topic and temporal information;

b) using Email corpus as data source for the ef-

fectiveness and feasibility test of the proposed framework;

c) discovering sentiment distribution and characteristics discovery in temporal categories and relationship between sentiment variance and topic categories.

2 Related Work

Sentiment analysis, a study of extracting and analyzing the implications of emotions, attitudes or opinions from natural language, has attracted researchers from diverse areas (Liu, 2015). Papers and articles on sentiment analysis published in recent years indicate a trend of more comprehensive view of conducting sentiment analysis, including feature enrichment, and sentiment visualization. Among them, research on sentiment with temporal or topic information is one of the most appealing targets.

As Liu (2015) illustrates in its definition of opinion, time is considered as a crucial factor in sentiment analysis as identifying pattern of sentiment changes from historical data assists in the trend prediction of the future, as well as the topic. Though some studies have been conducted on sentiment analysis with topic or temporal features, problems, such as limitations in the dataset options and no pattern display, remain unsolved (Diakopoulos et al., 2010; Fukuhara et al., 2007; Li and Liu, 2012; Mei et al., 2007). For example, Fukuhara et al. (2007) presented topic, timestamp and sentiment graph for news articles using coefficient model, even if the study was purely theoretical with insufficient experiments. Additionally, Mei et al. (2007) undertook experiments on discoverying relationship between topic and sentiment using topic-sentiment mixture model on webblogs; Diakopoulos et al. (2010) utilized Vox Civitas, an automated visual analytic tool, for extracting news from social media data stream, displaying topic and keyword trend. To the best of our knowledge, experiments using large Email data have not been proposed yet.

As for Email mining applications, reviews on previous articles reveal that information management and spam detection are heated study topics (Basavaraju and Prabhakar, 2010; Hangal et al., 2011; Tang et al., 2014; Whittaker and Sidner, 1996). For instance, Whittaker and Sidner (1996) highlighted the issue of Email overload and its negative influence on personal information management; Basavaraju and Prabhakar (2010) proposed a new approach for spam mail detection using semi-supervised learning algorithm. However, research on sentiment analysis using Email data is rare and leaves enormous space for refinement and improvement.

3 Framework

As illustrated in the previous section, a comprehensive and systematic framework is presented in this section. Figure 1 outlines major components and flow of the proposed scheme of approach. To be specific, the framework consists of several procedures, including data extraction, text preprocessing, feature selection and sentiment clustering.

Text preprocessing step incorporates basic Natural Language Processing (NLP) techniques, such as stop word removal and stemming. In feature selection process, topic, timestamp and opinion words are the feature options, in which topic is generated using keyword search and opinion words are generated using the English opinion lexicon (Liu et al., 2004). Sentiment clustering is composed of two substeps containing grouping data based on timestamp and classifying sentiment based on topic. DBSCAN algorithm is chosen to perform the clustering task for its efficiency in speed and effectiveness in handling noise (Ester et al., 1996).

3.1 Text Preprocessing

Text preprocessing aims at removing unnecessary information, such as punctuations and articles, and converting natural language into machine readable content. Tang et al. (2014) highlight the indispensability of Email data cleaning, whilst point out the complexity and limitation of the process. Herein, standard text preprocessing procedures for general NLP tasks have been applied.

In this study, Apache Lucene, an open-source NLP toolkit, is utilized for performing text normalization and filtering (Hatcher and Gospodnetic, 2004). Details are described as follows:

- First step: duplication removal and noise filtering. Assuming a dataset has been imported into

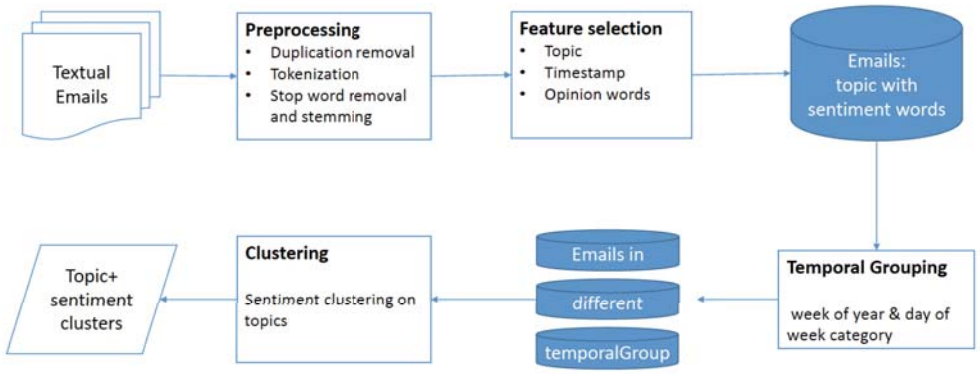

Figure 1: Framework for Email sentiment clustering in topic and temporal categories.

database, a query statement is required for retrieving the data. The implementation of SQL function *DISTINCT* and elimination of item with empty "subject" assist in the achievement of this process;

- Second step: tokenization. After retrieving the entire dataset from database, tokenizing each Email message through the implementation of *tokenize*() function for further processing;
- Third step: stop word removal and stemming. Filtering each Email message using *stopAnalyzer*() embedded in Apache toolkit removes common conjunctions, such as punctuation marks and articles, while iterating for stop word removal, simultaneously using *stem*() function for restoring words' original format, especially for verb and plural.

3.2 Feature Selection

Since the main aim of this research is to identify the sentiment distribution in accordance with topic and temporal classification, feature selection process is divided into three parts: topic word extraction, timestamp transformation and opinion words generation. Details of each feature category are discussed as follows.

Topic : feature extracted based on keyword search. A list of keywords matching topic category defined is generated manually. Querying column named *subject* in database returns a string containing subject data of each Email message. A comparison between the string and each keyword list is conducted for searching the corresponding category.

Timestamp : feature is generated through querying column named *data* in the database. Functions *getTime*() and *getTimeZone*() are implemented for converting temporal data into milliseconds with standard *UTC* timezone. For instance, date value "2016-04-02 11:12:28" is transformed into "1459559492612".

Opinion words : features identified on the basis of a well-defined English opinion lexicon (Liu et al., 2004) contain 2,046 positive words and phases, and 4,833 negative words and phases. Let $\mathcal{OW}$ be a collection of entire opinion lexicon, containing word $ow_1, ow_2, \ldots, ow_i$, then $\mathcal{OW} = \{ow_1, ow_2, \cdots, ow_i\}$ $i \in (6,879)$. Sample positive and negative words chosen from the lexicon are shown in Table 1.

Positive Words	Negative Words
good	bad
thank	disgrace
worthy	overwhelm
flourishing	incomplete
delight	sick

Table 1: Positive and negative words representation from the English opinion words list (Liu et al., 2004).

At this stage, a sequence of opinion words based on its presence in each Email message is stored for future study (i.e. the inner sentiment changes); however, bag-of-words model is adopted as a term weighting method for distance matrix for sentiment clustering, which will be discussed in the following

section. Each data item is transformed into feature representation after this procedure. A sample Email item is represented into:

$$< id, Topic, Timestamp, [ow_1, ow_2, ow_3, ow_4] >$$

3.3 Sentiment Clustering

Sentiment clustering is composed of two steps: grouping data based on timestamp and clustering sentiment based on topic. First step aims at clustering the entire dataset into different date categories with day and week labels using personalized Email temporal clustering algorithm. Second step clusters sentiment using DBSCAN clustering algorithm with bag-of-words term weighting method and Euclidean distance matrix in accordance with topic.

3.3.1 Grouping Data based on Timestamp

To investigate the Email distribution, an Email temporal clustering algorithm is applied to group Email messages into days under week category. The *Calendar* object embedded in Java is utilized for the comparison of timestamp with calendar and classification into day of the week. The pseudo code for *EmailTC* algorithm is presented in Algorithm 1. Due to the characteristics of *Calendar* object, the first day of week is defined as Sunday. Hence, the classification results start with day 1 representing Sunday and end with day 7 representing Saturday.

Algorithm 1 EmailTC

```
for each Email message e_i ∈ E do
    Get timestamp T from e_i
    Get Calendar object;
    Get week of year;
    Get day of week;
    Create group G_w for week of year;
    Create group G_d for day of week;
    if T ∉ G_w then
        Create subgroup Gsub_w;
        Put e_i in Gsub_w;
        Put e_i in G_d;
    else
        Put e_i in G_w;
        if T ∈ G_d then
            Put e_i in G_d;
        end if
    end if
end for
```

3.3.2 Clustering Sentiments based on Topic

Revised DBSCAN algorithm (Ester et al., 1996) with bag-of-words term weighting scheme (see Equation 1) and Euclidean distance (see Equation 2) as distance matrix are implemented for conducting the clustering process. Bag-of-words model and Euclidean distance, though invented for decades, remain efficient and well-adopted in many studies.

$$\mathcal{BOW} = freqeuncy * ow_i, i \in OW_p. \quad (1)$$

$$Eu_(d) = \sqrt{(x_s - x_t)^2 + (y_s - y_t)^2} s, t \in (1, n). \quad (2)$$

In Equation 1, supposing OW_p represents a collection of all positive opinion words, bag-of-words approach computes the frequency of each positive word ow_i appeared. Herein, Equation 2 calculates the distance between positive words and negative words contained in each Email message.

As for the option of DBSCAN, its ability of noise handling and fast processing speed increases the utilization of DBSCAN in various applications (Ester et al., 1996). Furthermore, as DBSCAN follows the rule of density-reachablility based on $minPts$ and $epsilon$ parameters defined, it generates diversified number of clusters in accordance with different sentiment scaling. The pseudo code for revised DBSCAN algorithm is presented in Algorithm 2.

Note that as Ester et al. (1996)'s DBSCAN algorithm served as the foundation of the revised version in this paper, more details can be referred to (Ester et al., 1996), especially for the $expandCluster()$ that has not been written out in the pseudo code due to its complexity. By changing the two parameters $minPts$ and $epsilon$, clustering results are varied accordingly (see Algorithm 2). Therefore, implementation of DBSCAN without accurate $minPts$ and $epsilon$ normally involves trial and error testing (Ester et al., 1996).

4 Empirical Results and Discussion

Experiments are conducted on a subcollection of the large Enron Email corpus, which contains emails exchanged from business operation, personal com-

Algorithm 2 AlgoDBSCAN

Input: A collection of Email messages $\mathcal{E}$, $minPts$, $epsilon$.
Output: A collection of sentiment clusters $\mathcal{C}$ with a subset of Email messages $\{E_1, \cdots, E_i\} \in (T_1, T_j)$.

/ Set $\mathcal{E}$ to UNCLASSIFIED*/*
for each Email message $e_i \in \mathcal{E}$ **do**
    Mark e_i as Cluster point c_i;
    Compute $BOW1$ for e_i;
    Compute $Eu_{(d)}$ between e_i and other data $\in \mathcal{E}$;
    Compare $epsilon$ with $Eu_{(d)}$ to find $\mathcal{N}$ neighbors;
    if $\mathcal{N}$ is greater than $minPts$ **then**
        Form cluster c_i;
        Insert e_i into c_i;
        Add all messages $\in \mathcal{E}$ reachable using expandCluster function;
    else
        Assign e_i to noise;
    end if
    Insert c_i into $\mathcal{C}$;
end for
return $\mathcal{C}$

munication, commercial and advertising. Graphs on the Email message distribution on temporal classification and sentiment distribution are topic classification are to be displayed for the visualization of the sentiment patterns discovered. In addition, sentiment words frequency is illustrated using tag cloud and frequency table.

4.1 Dataset

As Email data cleaning is an independent area requiring deep learning and investigation (Tang et al., 2014), a database version of the Enron Email corpus generated by (Liu and Lee, 2015) (available at $http://www.ahschulz.de/enron-Email-data/enron-mysqldump_v5.sql.gz$) has been utilized. A collection of $32,716$ Email messages exchanged between January to May in 2001 has been extracted from the Enron corpus database for conducting our experiments. MySQL database and Eclipse IDE are incorporated for data extraction and feature selection. *15* topic phases, such as $BusinessDocument$, $GeneralOperation$ and etc., are set up manually for grouping the dataset into different categories. Among them, a special topic named $Other$ is defined for storing messages with no subject keyword matching. As for temporal feature, the entire dataset is classified into *22* weeks with each subdivided into *7* days. All features are extracted using method discussed in the previous sections. The structure of data with feature representation is indicated in the following sample fragments (see Figure 2).

<3038, Company Strategy, 984614400000, [available, pretty, works]>
<3039, Technical Issue, 984614400000, [benefits, appreciate, good, -bad, good]>
<3041, Other, 984614400000, [comprehensive, convenient, concise, -issues, -concerns, available, dedicated, available]>

Figure 2: Fragments of data with feature representations.

4.2 Email Distribution in Temporal Categories

Experiments on sentiment clustering are undertaken using DBSCAN algorithm that requires two parameters $minPts$ and $epsilon$. With reasonable assumption and several attempts, the results are generated with $minPts$ of 5 and $epsilon$ of 0.15. Therefore, the description of sentiment clusters is assumed to be similar to a 5 likert scale including *strongly positive, positive, neutral, negative and strongly negative*. The graphs and tables of detailed sentiment clustering results are to be displayed in the following section. As a temporal clustering is performed before sentiment clustering, Fig. 3 and Fig. 4 illustrate the distribution of Email messages and clustering results in temporal categories.

In the two figures, 5 months are divided into 22 weeks with each week having 7 days. Based on the results shown in Fig. 3 and Fig. 4, more Emails exchange between weekdays with an average of 202 mails, than weekends with an average of 23 mails. This result is coherent with common observation that proves the authenticity of the dataset. Furthermore, in Fig. 4, more clusters are discovered during weekdays that implies a variety of topic discussed during business days. More detailed analysis on sentiment clustering results is to be discussed in the fol-

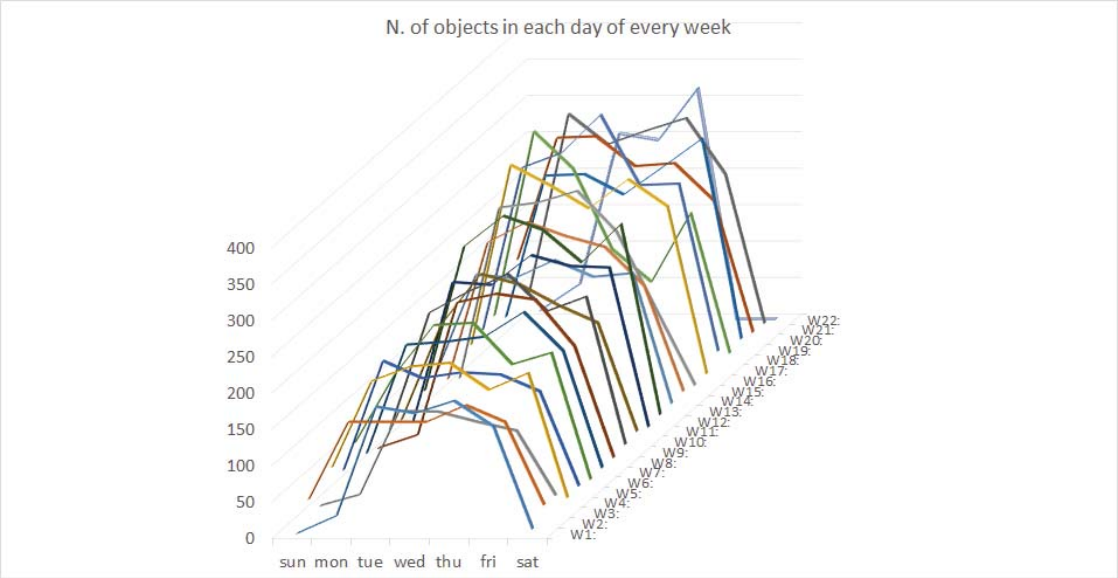

Figure 3: The distribution of Email messages in temporal category.

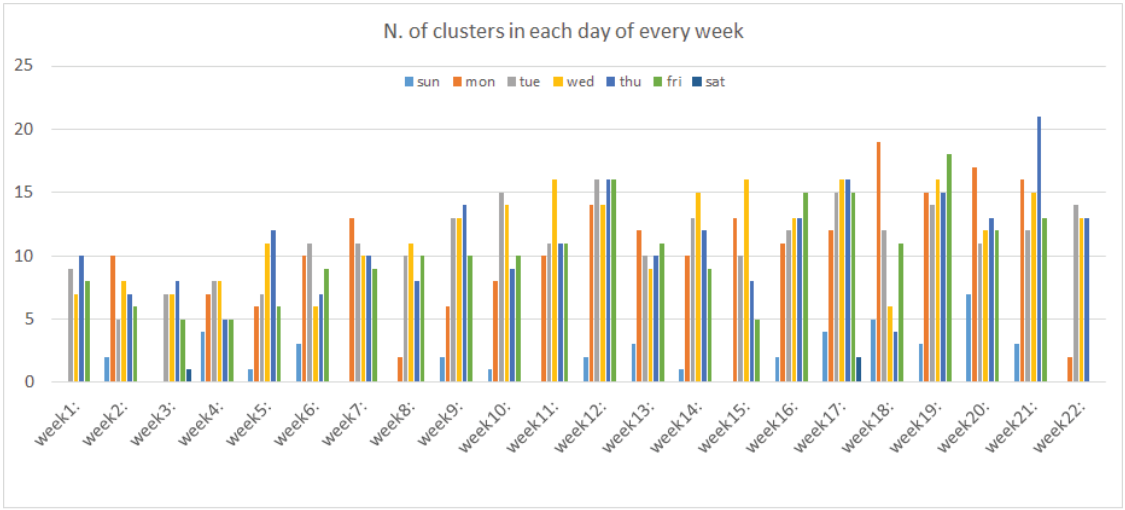

Figure 4: The distribution of Email clusters in temporal category.

lowing section.

4.3 Sentiment and Topic Clustering Results

As discussed in the previous section, sentiments are broadly categorized into five scales in accordance with the ratio of positive words and negative words. Since the research mainly focuses on finding distribution and patterns among sentiment clusters with topic and temporal information, accuracy of clustering results is not evaluated. The following two tables, Table 2 and Table 3, are generated using part of clustering results in week 4.

Due to the limitation of paper length, the above Table 2 summarizes some of the prominent clustering results with topic feature in each day. SP, P, $Neutral$, N, and SN represent for strongly positive, positive, neutral, negative and strongly negative, respectively. Individual cluster shows one topic

Day	Topic in sentiment clusters
Monday	Other-SP Private Issue-SP Commercial/Advertising-SP Company Strategy-SP General Operation-P Logistic Issue-NEUTRAL Other-SN
Tuesday	Private Issue-SP Other-SP Company Strategy-SP Logistic Issue-SP Other-P General Operation-P Logistic Issue-NEUTRAL Other-SN
Wednesday	Employee Training-SP Business Investment-SP Company Strategy-SP Company Project-SP Logistic Issue-SP Other-P General Operation-P Other-SN
Thursday	General Operation-SP Other-SP Employment Arrangement-SP Other-N Other-SN
Friday	General Operation-SP General Operation-P News/Press/Media-NEUTRAL Other-NEUTRAL Other-SN

Table 2: Sentiment clustering results in topic category in week 4.

with sentiments in that day. It appears that more positive clusters are discovered with various topics than negative clusters. Interestingly, some topics have both positive and negative clusters which indicates people's different views on the same topic that is coherent with human nature. Table 3 shows the corresponding items in some of the cluster.

The combination of two tables assists in the further justification of the option of DBSCAN input parameters and the sentiment result criteria. On one hand, objects are relevant in the corresponding cluster, while distinguished from others. For instance, Emails with more positive features are categorized into positive clusters, such as message id 73677 and message id 54522; while Emails with more negative features are categorized into negative clusters, such as message id 141463 and message id 180199. On the other hand, objects with different feature words are categorized into one cluster indicating a reasonable option of the $minPts$ parameter.

As an auxiliary to view the sentiments in details rather than a 5 likert scale, two tag cloud graphs (see Fig. 5 and Fig. 6) containing 100 positive and negative words and a table with 20 most frequently referred opinion words are displayed.

Positive	Frequency	Negative	Frequency
work	6767	issue	6896
support	3634	problem	3011
master	2448	limited	2865
thank	2207	risk	2818
lead	1762	crisis	2250
important	1489	concerns	2204
privileged	1277	vice	1763
respect	1262	error	1761
recommend	1220	debt	1410
helpful	1083	critical	1273

Table 4: Top 10 frequent opinion words.

Figure 5: Tag cloud for positive opinion words.

As shown in Table 4, positive words commonly used in Emails are work, support and master and negative words are issue, problem and limited. An applealing observation lies in the statistics that most

Clusters	Details
Cluster 1: (Other-SP)	⟨9948:Other-[like]⟩ ⟨24718:Other-[clear]⟩ ⟨34752:Other-[thank]⟩ ⟨47679:Other-[free]⟩ ⟨73677:Other-[available, good, significant]⟩ ⟨80082:Other-[love]⟩ ⟨103142:Other-[excel]⟩ ⟨103148:Other-[diligence, soft]⟩
Cluster 2: (Employment Arrangement-SP)	⟨24721:Employment Arrangement-[like, thank]⟩ ⟨33181:Employment Arrangement-[well, happy]⟩ ⟨54522:Employment Arrangement-[encouragement]⟩ ⟨80114:Employment Arrangement-[well, good]⟩ ⟨180206:Employment Arrangement-[works]⟩
Cluster 3: (Other-SN)	⟨120803:Other-[-curt]⟩ ⟨121264:Other-[-curt]⟩ ⟨141463:Other-[-miss, -miss, -miss]⟩ ⟨164561:Other-[-liars]⟩ ⟨164563:Other-[-hells]⟩ ⟨180199:Other-[-hedge, -issues, -hedge, -subjected]⟩

Table 3: Objects in sentiment clusters on topic category in week 4.

Figure 6: Tag cloud for negative opinion words.

negative words are nouns, while most positive words adjectives. Conclusion is to be summarized until more findings are discovered. However, a potential topic is developed for future study.

5 Conclusion and Future Work

In this research paper, we propose a robust and comprehensive framework for sentiment clustering with topic and temporal features using bag-of-words model as term weighting approach and revised DBSCAN algorithm for clustering. Features composed of topic, timestamp and opinion words are extracted for performing two experiments, including grouping Email messages into temporal categories and clustering sentiments based on topic, for discovering sentiment patterns.

Based on the summary of the clustering results, major findings are categorized into three points. First, the distribution of Email messages reveals more intense communication during weekdays and variety of topics. This pattern justifies the authenticity and originality of the Email corpus. Second, sentiment clustering results assist in the validation of the parameters chosen for the implementation of revised DBSCAN algorithm as clusters are distinguished from others as well as coherent within themselves. Also, feature words for generating sentiment clusters are with high similarity which implies a proper choice of the $epsilon$ value. And finally, Table 4 lists the top 20 most commonly used opinion words with most of them related to business and management, which shows a linkage to the characteristics of the dataset.

References

M. Basavaraju and Dr. R. Prabhakar. 2010. A novel method of spam mail detection using text based clustering approach. *International Journal of Computer Applications*, 5(4):15–25.

Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 115–122. IEEE.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. 2007. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *ICWSM*.

Sudheendra Hangal, Monica S Lam, and Jeffrey Heer. 2011. Muse: reviving memories using email archives. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 75–84. ACM.

Ming C Hao, Christian Rohrdantz, Halldor Janetzko, Daniel A Keim, Umeshwar Dayal, Lars erik Haug, Meichun Hsu, and Florian Stoffel. 2013. Visual sentiment analysis of customer feedback streams using geo-temporal term associations. *Information Visualization*, page 1473871613481691.

Erik Hatcher and Otis Gospodnetic. 2004. Lucene in action.

Gang Li and Fei Liu. 2012. Application of a clustering method on sentiment analysis. *Journal of Information Science*, 38(2):127–139.

Sisi Liu and Ickjai Lee. 2015. A hybrid sentiment analysis framework for large email data. In *Intelligent Systems and Knowledge Engineering (ISKE), 2015 10th International Conference on*, pages 324–330. IEEE.

Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2004. Text classification by labeling words. In *AAAI*, volume 4, pages 425–430.

Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

Alvaro Ortigosa, José M Martín, and Rosa M Carro. 2014. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46.

Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012*, pages 508–524. Springer.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2013. Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88. Springer.

Guanting Tang, Jian Pei, and Wo-Shun Luk. 2014. Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems*, 41(1):1–31.

Steve Whittaker and Candace Sidner. 1996. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 276–283. ACM.