# Sentiment analysis of text with lossless mining

Abdul Razaq
Sanaz Kavianpour
Gavin Hales

Razaq, A., Kavianpour, S. & Hales, G. (2021) 'Sentiment analysis of text with lossless mining'. In: *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2021)*. IEEE , Piscataway, NJ, International Conference on Electrical, Computer, Communications and Mechatronics Engineering, Mauritius, 7-8 October 2021. DOI: https://doi.org/10.1109/ICECCME52200.2021.9591155

# Sentiment Analysis of Text with Lossless Mining


Abdul Razaq
School of Design and Informatics
Abertay University
Dundee, Scotland, UK
a.razaq@abertay.ac.uk

Sanaz Kavianpour
School of Design and Informatics
Abertay University
Dundee, Scotland, UK
s.kavianpour@abertay.ac.uk

Gavin Hales
School of Design and Informatics
Abertay University
Dundee, Scotland, UK
gavin.hales@abertay.ac.uk



***Abstract*— Social networks are becoming more and more real with their power to influence public opinions, election outcomes, or the creation of an artificial surge in demand or supply. The continuous stream of information is valuable, but it comes with a big data problem. The question is how to mine social text at a large scale and execute machine learning algorithms to create predictive models or historical views of previous trends. This paper introduces a cyber dictionary for every user, which contains only words used in tweets – as a case study. Then, it mines all the known and unknown words by their frequency, which provides the analytic capability to run a multi-level classifier.**



***Keywords*— *lossless mining, sentiment analysis, social media analytics, social network***


## I. INTRODUCTION

The information universe has been on the ascent for the last few years, and as a result, unstructured information in cyberspace has likewise expanded. One of the essential supporters of this unstructured information is online networking sites such as Facebook, LinkedIn, Twitter, Tencent QQ, Google , etc [1]. These online networking sites enable individuals to share and express their perspectives about subjects, have a dialogue with various groups, or post messages. For example, Twitter is an extremely prevalent online networking site, where roughly 6000 tweets are sent every second per day by around 500 million users.

The massive amount of data generated via the individual's activities is extremely rich in content and can be used for data analytics. Application of data mining techniques on data can assist in finding hidden patterns from data and learn many indictors. Sentiment analysis can be used in various levels, including sentence, aspect, document, and user level and can be done via Machine Learning (ML) techniques such as clustering or classification, lexicon, etc. An analytical method is a practice of relating to or using analysis or logical reasoning. In terms of Artificial Intelligence (AI)/Machine learning (ML), it is a technique to interpret data into a meaningful format [2]. The fundamentals of data analytics are very much based on common practices of statistical and mathematical theories.

There has been a growing body of work in the field of opinion investigation with social platforms. Advertisers and firms utilize this examined information to seek the current trend, create predilection models, or view historical events [3]. With the enormous amount of data on the web from various informal communities such as Twitter, Facebook, WeChat, etc., programmes that can figure out what individuals are discussing in specific areas and over a certain timeframe are required. A considerable number of existing programmes have accomplished excellent results regarding identifying the subject topic, essentially gathering comparable information together, demonstrating that they are about a similar point. Point extraction can be viewed as a subsequent stage, where the objective is to mark these gatherings by separating a subject title for each information gathering. In the current era of social media platforms, an enormous measure of information comprises genuine beliefs, or real-time events are disseminated and consumed consistently. This paper examines the different machine learning algorithms utilized in Twitter information streams - as a case study - and presents a text mining analytics system for sentiment analysis. The proposed scheme is a system that deploys rule-based feature mining for all words with a machine learning classifier to construct categories. These categories can reveal specific preferences, occasions, brands etc. through sentiment analysis.

The remainder of the paper is arranged as follows. Section II explores the current literature base to evaluate existing techniques of sentiment classification and analysis systems. In section III, the proposed scheme is presented followed by details of the system design, the framework, and the testbed and implementation methodology as well as presenting the initial results. Section IV presents the conclusions from the paper.

## II. RELATED WORK

Social media is a robust and low-cost information exchange platform, enabling ideas to circulate amongst various parties such as its consumers, organizations, or any other entity to learn, share, build, market, advertise, etc. Its significant revenue is the generation of vast amounts of information that presents users with outstanding service [4]. The generated vast amount of information is an indispensable source of big data associated with sentiment or the opinion sphere. Through sentiment analysis, meaningful patterns can be extracted from texts and are available to various online sources. For instance, companies can detect the popularity of their products and use their findings to predict the future of their production. Some of the related work

that has been done on social media mining is presented as follows.

Dokoohaki et al. [5] investigated Twitter messages for European and General elections. They used PageRank of political tweets utilizing a stochastic connection structure mining approach, to compare the vote result and Twitter trends. They confirmed that Twitter information could be utilised to foresee the result of the political races. Kharde and Sonawane [6] presented an overview and relative investigation of existing techniques for opinion mining including machine learning and lexicon-based methodologies. There are two main approaches for Sentiment Analysis (SA) as corpus-based and lexicon-based. The lexicon-based approach entails calculating the sentiment from the semantic orientation of word or phrases that appear in a text. A dictionary of positive and negative words is required for this approach including a positive or negative sentiment value given to each word. Furthermore, the local context of a word including negation or intensification also needs to be regarded. The final prediction is based on a merging function like sum or average, concerning the total sentiment for the message. The corpus-based approach has a goal of giving word references identified with a particular domain. It includes exploiting co-incidence statistics or syntactic patterns in a text corpus. These lexicons are created from an arrangement of seed conclusion terms that evolve through the inquiry of related words, using either factual or semantic strategies. Kharde and Sonawane employed various ML algorithms such as Naive Bayes, and Support Vector Machine (SVM) on Twitter data streams, and highlighted the challenges of Sentiment Analysis application on Twitter. Based on the achieved results of their research, ML methods delineate high accuracy and can be considered as the baseline learning methods. On the other hand, lexicon-based methods show effectiveness in some cases with minimum effort human-labelled documents. Hence, integration of ML methods to augment lexicon methods can advance the accuracy of sentiment classification and adaptive capacity of both various domains and languages.

Sharma et al. [7] gathered vast Twitter messages for eight months, made by 79,768 Twitter clients, and channelled them through Natural Language Processing (NLP) to create different categories. Rafea and Mostafa [8] investigated Arabic text with k-means. The frequent words (words that occurred more than 20 times) in the entire corpus were utilized as elements to arrange the tweets via the k-mean classifier. Most recently, Satyen M. et al. utilized Python-based Tweepy solution to extract Twitter content to execute the classification algorithm. The features are extracted and classified - among positive, negative and neutral - using N-gram modelling technique with supervised ML algorithms of SVM and KNN (K-Nearest Neighbor) for classification [9].

Joshi and Tekchandani [10] extracted data (e.g., movie reviews) from Twitter and applied SVM to calculate maximum entropy and Naïve Bayes to classify data using unigram, bigram, and hybrid. The results delineated that SVM outperforms other classifiers with an accuracy of 84% for movie reviews. Romsaiyud et al. [11] enhanced the Naïve Bayes classifier to extract the words from Twitter and examine loaded pattern clustering. Their aim was to prevent cyber victimization from Cyberbullying. A cyberbully can be characterized as an individual who utilizes electronic types of correspondence to post or send content to an individual that they would consider debilitating, humiliating, or badgering. The most typical types of cyberbully are sending somebody's private messages to others, sending debilitating or improper messages using texting, and hacking into someone else's record to send unacceptable data to others through informal community destinations. Their developed algorithms employed two main methods as follows. First, Using k-mean clustering to create partitions by iteratively relocating from entire datasets into clusters and second, gaining any specific partition with the frequency of words by multinomial model feature vector, and illustrating the probability of words appearing in a document for predicting the classes. Their proposed strategy produced a prescient model from an extensive volume of informational indexes for supporting the investigation of benefits in business. The strategy was executed on CyberCrime Information, a physically marked dataset for 170,019 posts, and a Twitter site for 467 million Twitter posts.

Barnwal et al. [12] examined 15K + Tweets utilizing a strategy to study the significant inquiries concerning Indian Healthcare. They investigated via their trial setup how the online social networking Twitter can be used to identify the top subjects being talked about. Hao et al. [13] utilized Word2vec and Dynamic Conditional Random Field (DCRF) based system for Assumption Examination of Microblog WeChat. As challenges of the Microblog message, for example, the length and vocabulary restrictions, Word2vec innovation is utilized. In this technique, they proposed to use Word2vec to expand each WeChat post; At that point, DCRF display is utilized to consolidate frequent favourable words. Srivatsa et al. [14] examined the real nature of semantic abundance to manage the generous volume of text reports, which frequently don't have unequivocal metadata qualities. The volume of text can push bottleneck resources to the limit. They proposed to model and measure semantic overabundance in immense volume content streams using subject models and conclusion examination.

## III. PROPOSED SCHEME

Let $T$ be a collection of Twitter users $u$ with $p$ profile and messages $m$. A profile consists of user metadata (screen name, location, time zone, device type, connection type, etc.), and the Message is content generated by a user. Every message has a limitation of 280 characters, and their occurrence can vary from user-to-user. A message consists of sentence $s$; a sentence is constructed with words $w$.

Message content is mined in the form of words based on supervised data from various resources to build a Dictionary for each user. A word is checked against Merriam-Webster or emoji, and if it does not return a valid response, then the word is tagged with unknown words to simplify proof-of-concept. This limitation can be improved in conjunction with work conducted by IBM on the Watson platform. They semantically mine Twitter messages to correct the word or sentence. A similar approach adopted by auto-correct techniques can increase the system scope.

System Parameters:

$w = \{unknown, alphabet, number, merriam - webster, emoji\}$

Collection of users

$T = \{u_0, u_1, u_2, \ldots, u_N\}$

Messages of individual users $u_i = \{m_0, m_1, m_2, \ldots, m_N\}$

Sentences in a tweet $m_i = \{s_0, s_1, s_2, \ldots, s_N\}$

Words in sentences $s_i = \{w_0, w_1, w_2, \ldots, w_N\}$

$observation : o = w$

### A. Message Anatomy

Messages or so-called tweets can be divided into different components. A message can be a Hashtag, country names, companies, or plain construction of words. The maximum limit of each message is 280 characters, but it can include URL, images, etc. This leaves an opportunity to integrate Optical Character Recognition (OCR) or Image Analysis. Table $T$ is a matrix with columns as observations and rows as tweets. A single tweet is connected to user profile ($P$) Table, which includes some static and dynamic data. A Twitter handle is considered unique and non-transferable, so the Twitter handle is used as a unique identifier in Table $P$ mapped to Tables I and II. Profile Table includes meta-data: The Twitter handle is mapped to Unique identifier ($uid$) and is set to constant, whereas the rest of the fields is updated with a timestamp including screen name, location, time zone, device type, connection information, browser information. A tweet is parsed and mined with lossless compression. If the user id of a new tweet does not exist in Table $T$, then a new entry is created in profile Table $P$. The content of a new message is parsed with a string tokenizer to create an array of independent words as illustrated in the 'Create list of words' process of Fig. 1 flowchart. The words are associated with the user if they exist in the Dictionary, and in case of unknown strings, they are added to the user collection as a new record. The idea is to mine all the known and unknown words and provide lossless data preservation.

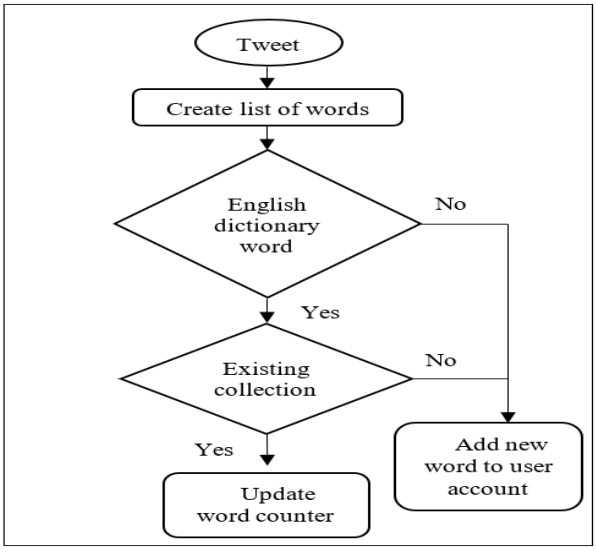


Fig.1. Flowchart-mining a tweet.

### B. Mining a Tweet

The messages are mined based on their content. If there is an existing corresponding word already in the user dictionary, then the counter of value is increased to reflect the frequency. If the word does not exist in the user dictionary, then a new column for a word is added, and the counter value is set to one.

Table I presents the same tweets for a user with the User ID of $User01$, and three tweets are shown with a timestamp with three different greetings. All the messages are mined in Table II. The first message of 'Good Morning' is mined in the first row with a counter value set to 1. However, the second message contains a repetition of the word Good, which is mined with the previous tweet. Thus, the second row is added with a counter set to 2 for the word Good and a new column for Afternoon. The third tweet is also mined with a similar method.

The pseudo-code in Algorithm 1 presents the blueprint of the message miner with three sources of references, for example, Merriam-Webster, Oxford, and emoji. All incoming words from user tweets are searched through these indexes, and a match is identified upon successful search. In the case of unknown, the miner will all the word, but it will flag with $unknown$ tag.

TABLE I. MESSAGE CONTENT

| **User01 Tweets** | |
|---|---|
| ***Time*** | ***Content*** |
| August 26 08:45:30 2020 | Good Morning |
| August 26 09:30:40 2020 | Good Afternoon |
| August 26 09:35:50 2020 | Good Evening |

TABLE II. MINED MESSAGES

| **User01 Features** | | | | | |
|---|---|---|---|---|---|
| ***Time*** | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_N$ |
| August 26 08:45:30 2020 | 1 | 1 | | | |
| August 26 09:30:40 2020 | 2 | | 1 | | |
| August 26 09:35:50 2020 | 3 | | | 1 | |

Table III illustrates the basic dictionary of a user if all the tweets are mined in the proposed method over time. The Dictionary will contain all the words written by a user. Effectively each user will have their dictionary consisting of words that are written by them. And Table IV includes user profile settings. This table can be expended to include more data such as device type, battery level, IP etc.

TABLE III. LOSSLESS QUANTITATIVE COMPRESSION

| **User Dictionary** | |
|---|---|
| ***index*** | ***word*** |
| $o_1$ | Good |
| $o_2$ | Morning |
| $o_3$ | Afternoon |
| $o_4$ | Evening |
| $o_N$ | |

TABLE IV. USER PROFILES

| Unique Records with Dynamic Status | | | | |
|---|---|---|---|---|
| *UID* | *friends* | *followers* | *location* | *lastUpdate* |
| user01 | 34 | 120 | UK | August 01 01:12:30 2020 |
| user01 | 34 | 120 | UK | August 26 09:30:40 2020 |
| user01 | 35 | 121 | UK | August 26 09:55:50 2020 |

### C. Data Classifier

In this section, a generic framework of application is discussed to provide control and tuning options (input filters). These input options provide a way to create and execute different hypotheses based on the particular use. K-mean clustering is proposed to perform the analytics. A filter based on a set of words can serve as cluster boundaries. For example, to analyse user movies preference, a filter based on imdb.com can construct a multi-classifier to show user preference.

**Algorithm 1 Message Mining**

```
int Dictionaries (int word) {
     if (Merriam −Webster (word))
              return 1;
     if (oxford (word))
              return 1;
     if (emoji (word))
              return  1;
/* Add various dictionaries to expand */
return   0;
}
   int       parsepayload (void) {
   int   Observation, USER_DICTIONARY,
         Max_MSG_SIZE, ObservationCount;

   char o [Observation] [USER_DICTIONARY];
   char w [ Max_MSG_SIZE];
/* USER_DICTIONARY= 0 for first dictionary */
/* ObservationCount = The number of words in a message*/
    for (int i=0; i < Max_MSG_SIZE; i ++) {

        if (w [i] == o [i] [0])
            o [i] [0] + + ; /* Word exist in User's dictionary */
        else if (Dictionaries (w[i]) /* New word detected*/
            o [ObservationCount++] [0] = w[i];
        else  /* New user defined word */
            o [UnknownCount++] [0]= w[i];
   }
      return SUCCESS;
}
```

A k-mean clustering for four clusters can analyse user preference based on Table II and IV. The first column of Table V contains four types of categories with corresponding 'words' to define the genre. Table III includes the collection of words that were aggregated over time. A hypothetical user behaviour as presented in Fig. 2 can be built from Table III that is mapped to the genre mapping in Table V. The user behaviour can be built in more detail while expanding Table V to include more preferences. This proposed expansion can identify psychological trends over time and can be used for various purposes such as security, monitoring, customized services, or target marketing.

TABLE V. FILTER CONSTRUCTION

| Genre Mapping | |
|---|---|
| Genre | User Words In K-mean Clusters |
| Drama | Emotion, discussion, argument, theatre, narrative,serious, in-depth, emotional, struggles |
| Fantasy | Magic, supernatural, forces, technology, fairy, tales, legends |
| Crime | Bad, Courtroom, police, detective, lawyer, judge, courts, legal, jurisdiction, thief, robber, fraud |
| Comedy | Good, funny, comical, jokes, laugh, satire, humour, parody sarcasm, stereotyping, mockery |

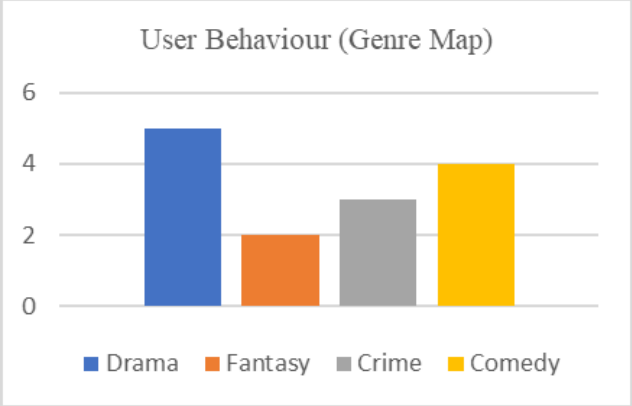


Fig.2. User behaviour.

## IV. CONCLUSION

The application scope is to mine a vast sum of data to find a pattern at the population level. The continuous data gathering is intrusive, but it is so significant not to anonymize the data completely. An application should anonymize the data and keep the mapping to identify the user-for possible observation if required. The message content and user identities can be cross-referenced with pseudo values, and the mapping information should be kept away from Admin or Users. It ensures that applications are not invading personal boundaries.

The method adopted to mine messages does not include semantic sense as tweets are not saved as they are; however, the semantic analysis can be applied from mined data proposed here. The problem of saving original tweets to preserve sentence sense quickly raises data storage issues.

The proposed solution stores full message content without discarding any information with lossless mining. It provides the ability to implement more complex and advanced machine

learning algorithms to be implemented over a vast set of features. It also provides an efficient and novel way to store data without vast storage requirements.

## REFERENCES


[1] I.Papasolomou and Y.Melanthiou, "Social media: Marketing public relations'new best friend," Journal of Promotion Management, 18(3), 2012, pp. 319-328, doi: 10.1080/10496491.2012.696458.

[2] W. Akinfaderin, "The Mathematics of Machine Learning," 2017.
Available at: https://towardsdatascience.com/the-mathematics-of-machine-learning-894f046c568

[3] C.W. Phang, A. Kankanhalli, and B.C. Tan, "What Motivates Contributors vs. Lurkers? An Investigation of Online Feedback Forums," Information Systems Research, 26(4), 2015, pp.773-792, doi: 10.1287/isre.2015.0599.

[4] K.K. Kapoor, K. Tamilmani, N.P. Rana, P. Patil, Y.K. Dwivedi, and S. Nerur, "Advances in Social Media Research: Past, Present and Future, " Inf Syst Front 20, 2017, pp. 531-558.

[5] N. Dokoohaki, F. Zikou, D. Gillblad, and M. Matskin, "Predicting swedish elections with twitter: A case for stochastic link structure analysis, " In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug 2015, pp. 1269-1276.

[6] V.A. Kharde and S. Sonawane, "Sentiment analysis of twitter data: a survey of techniques, " International Journal of Computer Applications 139(11), 2016, pp.5-15, doi:10.5120/ijca2016908625.

[7] A. Sharma, T. Yuan, and D. Lo, "What's hot in software engineering twitter space?, "In 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME), Sept 2015, pp. 541-545.

[8] A. Rafea and N.A. Mostafa Gaballah, "Topic extraction in social media, "In 2013 International Conference on Collaboration Technologies and Systems (CTS), May 2013, pp. 94-98, doi: 10.1109/CTS.2013.6567212.

[9] S.M. Parikh and M.K. Shah, "Classification of Sentiment Analysis Using Machine Learning, "Innovative Data Communication Technologies and Application (ICIDCA 2019), vol. 46, Springer.

[10] R. Joshi, and R. Tekchandani, "Comparative analysis of twitter data using supervised classifiers, " In 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 3, Aug 2016, pp. 1-6.

[11] W. Romsaiyud, K. Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns, " In 2017 9th International Conference on Knowledge and smart Technology (KST), Feb 2017, pp.242-247, doi: 10.1109/KST.2017.7886127.

[12] A.K. Bamwal, G.K. Choudhary, R. Swamim, A. Kedia, S. Goswami, and A.K. Das, "Application of twitter in health care sector for india, " In 3rd International Conference on Recent Advances in Information Technology (RAIT), March 2016, pp. 172-176, doi: 10.1109/RAIT.2016.7507896.

[13] Z. Hao, R. Cai, Y. Yang, W. Wen, and L. Liang, "A dynamic conditional random field-based framework for sentence-level sentiment analysis of chinese microblog, "In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, July 2017, pp. 135-142, doi: 10.1109/CSE-EUC.2017.33.

[14] M. Srivatsa, S. Lee, and T. Abdelzaher, "Mining diverse opinions, " In MILCOM 2012 - 2012 IEEE Military Communications Conference, Oct 2012, pp. 1-7, doi:10.1109/MILCOM.2012.6415602.