Social Network Analysis and Mining (2021) 11:24
https://doi.org/10.1007/s13278-021-00728-0

REVIEW PAPER-PRODUCTION ENGINEERING

National happiness index monitoring using Twitter for bilanguages

Di Wang[1] · Ahmad Al-Rubaie[1] · Benjamin Hirsch[1] · Gregory Cameron Pole[2,3]

Received: 21 September 2020 / Revised: 31 December 2020 / Accepted: 23 January 2021

Abstract
Nowadays, social media have become one of the most important methods of communication that provide a real-time and rich source of information, including sentiments. Understanding the population sentiment is a key goal for organisations and governments. In recent years, quite a lot of research has been done on sentiment analysis from social media. However, all the work in the state of the art is focused on a specific pre-defined subset of tweets, e.g. sentiment analysis via keywords search from tweets for relevant brands, products, services, events and so forth. Monitoring the general sentiment at national level through the whole social media stream is not done due to the challenges of filtering sentiment-irrelevant information, diversity of vocabulary usage in general tweets across topics causing low accuracy and the need for bilingual or multilingual models. This paper proposes a system for general population sentiment monitoring from a social media stream (Twitter), through comprehensive multi-level filters, and our proposed improved latent Dirichlet allocation (LDA) (Wang et al. in ACM Trans Internet Technol 18(1):1–23, 2017; Wang and Al-Rubaie in Appl Soft Comput 33:250–262, 2015; https://patents.google.com/patent/US20170293597A1/en) method for sentiment classification. Experiments show that our proposed improved LDA for sentiment analysis yields the best results, and also validate our proposed system for national sentiment monitoring in Abu Dhabi using twitter.

Keywords Twitter analysis · Sentiment analysis · Latent Dirichlet allocation (LDA) · Happiness index · Short message classification · Bilanguage model

Introduction

Social media are now recognised as an important source of near-instantaneous information on events, news, ideas and more importantly opinions and emotions. It represents an effective and immediate method of communication among individuals and communities. The informal nature of social media has enabled it to become a rich medium for direct exchange of opinions and expressing sentiments, hence an effective medium for carrying out sentiment analysis to monitor happiness indices close to real time.

To understand the general sentiment of a certain populace is an important goal of organisations and governments globally. Social media offer first-hand insight into the thoughts, feelings and concerns of the population. Therefore, monitoring social media enables organisations and government to measure the degree of population happiness without involving significant costs.

Some research has been carried out on sentiment analysis from social media, but most of which focuses on sentiment analysis for a particular service or product/brand and mainly for languages with simple grammar, such as English. However, general sentiment analysis from short messages (tweets) for multiple languages has been challenging due to the following reasons:

1. Limited information in very short messages—the use of short text with informal expressions and word variations caused by spelling errors, tweet slang and abbreviations,
2. Huge amount of noisy data—we aim to obtain as many tweets as we can access within a region/country, and therefore, we need to filter out tweets that are irrelevant

✉ Di Wang
di.wang@ku.ac.ae

Gregory Cameron Pole
gregpole@hotmail.com

[1] EBTIC, Khalifa University, Abu Dhabi, UAE

[2] Statistic Centre Abu Dhabi, Abu Dhabi, UAE

[3] Federal Competitiveness and Statistics Authority, Dubai, UAE

Published online: 26 February 2021

to populace sentiment. This is very challenging, considering that many organisations and entities use Twitter for their marketing activities, events/news/announcements, adding to advertisements, spam and many other sentiment-irrelevant tweets.

3. Sentiment is subjective—different persons might have different understanding and interpretation of the same tweet in the context of sentiment expression. Hence, sentiment analysis from tweets is even more challenging than other tweet applications, e.g. topics classification from tweets.
4. General sentiment analysis is more challenging—general sentiment analysis is more difficult than that for particular services/products or events, as it needs to cover a huge amount of vocabulary to represent sentiment across all areas.
5. Challenge of sentiment analysis from tweets in the Arab world—many tweets in the Arab world are written in Arabic or a mixture of Arabic and another language such as English. Text mining techniques are mature to deal with commonly used and relatively simple grammar languages such as English. Techniques and methodologies for Arabic text mining (especially for short messages with informal expressions) are still immature due to the inherent complexity of the Arabic language in terms of both structure and morphology. Hence, the proposed method and system have to be language independent and able to deal with different languages, including complex ones, while still achieving acceptable accuracy.

However, due to the wide usage of social media in Arabic countries, there is great demand for an accurate system to monitor the happiness index of general sentiment from social media. Such a system needs to be at least bilingual. Given the challenges and high demand above, this paper proposes a general system for population happiness index monitoring using sentiment analysis from a social media stream (Twitter). The contributions of this work can be highlighted as:

1. Obtain an accurate filter to identify sentiment-relevant tweets. We propose and use multi-level comprehensive filters. An accurate filter is the premise of an accurate sentiment analysis. The filtering includes (1) Bayes Nets filtering out sentiment-irrelevant tweets by features, e.g. length of tweets, whether they include emoji, mentions @, mobile numbers, etc.; (2) improved LDA classifier (applied on text) to classify tweets as sentiment-relevant tweets or non-sentiment tweets.
2. Achieve a better overall accuracy for general sentiment analysis. We applied our proposed improved LDA for tweets (Wang et al. 2017; Wang and Al-Rubaie 2015)[1] classification to achieve similar or better accuracy for sentiment analysis than human being tagging. The proposed method is proved to obtain better results on benchmark datasets than the state-of-the-art methods. The proposed method is language independent and can deal with both English and Arabic tweets.
3. Provide a comprehensive tweet sub-stream selection which can filter, then display sentiment per theme or focus subject matter, e.g. particular topic(s), time period, location(s) and specific event(s).
4. Develop a comprehensive tweet visualiser that is flexible enough to show the user sentiment either in general or for a particular topic of interest(s), as time series, geo-location distributions and table of tweets.

This paper is organised as follows: we discuss relevant work to sentiment analysis from social media in Sect. 2 and present our proposed system in Sect. 3. To prove the validity, we applied our proposed method to benchmark datasets and the results and comparisons are discussed in Sect. 4. Section 5 presents the results of our proposed method when used for general populace sentiment analysis in Abu Dhabi, UAE. At last, Sect. 6 discusses our conclusion and future work.

Related work

Diverse research has been carried out to investigate how to improve the accuracy of sentiment analysis from social media. Some research focussed on the importance of pre-processing for sentiment analysis from social media (Krouska et al. 2016; Singh and Kumari 2016; Sahu et al. 2015; Jianqiang and Xiaolin 2017). For example, Singh and Kumari (2016) investigated the importance of pre-processing and claim that their proposed method is robust to size of datasets but also improved the accuracy. Some research was done using ensemble methods (Troussas et al. 2016; Hassan et al. 2013; Kanakaraj and Guddeti 2015; Abdelwahab et al. 2015; Zhao 2016), removing the imbalance of dataset and solving the problem of not enough training data, with the aim of improving the overall accuracies. Hassan and Abbasi (2013) proposed a bootstrap ensemble framework for Twitter sentiment analysis to solve the data imbalance problem and achieved better accuracy (between 27 and 80% across different datasets). Kanakaraj and Guddeti (2015) analysed how different ensemble methods help with the accuracy improvement for sentiment analysis from twitter. Zhao (2016) used ensemble learning (representing words as vectors) and combined sematic and prior polarity for boosting

[1] https://patents.google.com/patent/US20170293597A1/en.

Twitter sentiment analysis. Krouska et al. (2017) presented a system for sentiment analysis for figurative language (emotion) based on both content (text) approach and an emotion pattern approach, which showed that combining context analysis and emotion pattern analysis improves the overall accuracy for sentiment analysis from social media. Krouska et al. (2016), Troussas et al. (2016), and Nguyen and Jung (2017) evaluated how the pre-processing (Krouska et al. 2016) and data ensemble (Troussas et al. 2016) help improve the sentiment analysis accuracy for different machine learning methods, and showed ensemble improves the accuracy significantly. They also compared the performance of using different algorithms for sentiment analysis on benchmark datasets (Nguyen and Jung 2017). Abdelwahab and Bahgat (Abdelwahab et al. 2015) investigated the relationship between the size of training set and accuracy. Sahu and Rout (Sahu et al. 2015) used a lexicon-based method for sentiment analysis and claimed that sentiment analysis in different domains need different lexicons, which indicates the inflexibility of lexicon-based methods. Kontopoulos et al. (2013) proposed an ontology-based sentiment analysis for tweets. Furini and Montangero (2016) proposed and investigated a gamification approach for sentiment analysis from tweets. Bravo-Marquez et al. (2016) investigated the transferring between words in sentiment lexicon and tweets' sentiments. From the state of the art, machine learning techniques are also applied to sentiment analysis including naïve Bayes, maximum entropy (MAXENT) and support vector machine (SVM), as well as artificial neural network (ANN). Nakov et al. (2016a, b) discussed different algorithms and performances for sentiment analysis from short messages based on benchmark datasets, SemEval. Their investigation shows that most systems were supervised and used a variety of handcrafted features and Twitter-specific encodings. Chiassi et al. (2013) and Zimbra et al. (2016) proposed hybrid system using n-gram and dynamic artificial neural network for brand sentiment analysis based on manually defined lexicon set and claimed their accuracy to be over 95%. Ajay and Sudhir (AIIT 2016) compared the performance of six popular machine learning methods for sentiment analysis from tweets. Shyamasundar and Jhansi (2016) focused on feature extraction using TF-IDF to improve the sentiment analysis accuracy from machine learning techniques (naïve Bayes and SVM) for benchmark datasets. Siddiqua et al. (2016) proposed a rule-based classifier based on the emoji and sentiment lexicons for Twitter sentiment analysis and compared with two known systems (SentiStrength and Semantria) and claimed better accuracy in three out of four benchmark datasets.

Some of the proposed methods are applied to various benchmarks or their own collected datasets and achieved accuracy between 60 and 85% depending on the algorithms and datasets. Ren et al. (2016) improved the word embedding method for Twitter sentiment analysis and achieved an accuracy of 78.57% without expert knowledge and 81.02% with expert knowledge for benchmark datasets. Wu et al (2016) proposed a method to extract sentiment knowledge from massive unlabelled data (tweets) and achieved an accuracy of 85% to 88% when applying it to different datasets. Saifa et al. (2016) introduced a lexicon-based approach which achieved better results for two datasets but worse result for the third. Pandey et al. (2017) proposed a novel *k*-mean-based cuckoo search for sentiment analysis from tweets which showed an accuracy of 67% to 84% for different benchmark datasets. Zhao and Cao (2015) proposed a method to make use of the co-occurrence statistics and contextual semantic relations as features for sentiment analysis and produced an average accuracy of 82% for different benchmark datasets. Christos and Maria tested different algorithms for sentiment analysis services. In this context, five well-known learning-based classifiers (naïve Bayes, support vector machine, *k*-nearest neighbour, logistic regression and C4.5) and a lexicon-based approach (SentiStrength) have been evaluated based on confusion matrices, using three different datasets (OMD, HCR and STS-Gold) and two test models (percentage split and cross-validation). The results demonstrated the superiority of naïve Bayes and support vector machine regardless of datasets and test methods.

Some methods have been applied to particular real-life problems, e.g. sentiment for a particular brand for customer services, events or political elections. Much research has been done for sentiment analysis from tweets for brands for customer services (Chiassi et al. 2013; Zimbra et al. 2016; Qaisi and Aljarah 2016). Philander and Zhong (2016) used a dictionary-based method for social media microblogging data sentiment analysis for hospitality operators and used Las Vegas integrated resort casino as use case, which they claimed to yield reasonable reliability for reviews ranking. Qaisi and Aljarah (2016) used tweet sentiment analysis for two different cloud service providers. Gupta and Kohli (2016) used Twitter sentiment analysis in healthcare. Yang and Wang (2014) proposed a method to analyse real-time tweets for US soccer fans during 2014 FIFA World Cup games and analysed the dynamic emotion changes of fans through the courses of the games. Schumaker et al. (2016) used sentiment analysis from social media to predict football matches outcomes and yielded an accuracy of 67%. Peng et al. (2016) used twitter sentiment analysis for drugs incidents using Weka packages and the sentiment analysis accuracy was just over 62–65%. Ramteke et al. (2016) used Twitter sentiment analysis for election result prediction. Other applications included how to use social media to assist in solving other problems, e.g. economic analysis and financial prediction by using social media as supplementary information. Smailovic et al. (2014) proposed a system to use social media to help financial markets prediction for more

profitable online trading. Porshnev et al. (2013) used different dictionaries for sentiment analysis from tweets to help financial and stock prediction. Daniel et al. (2017) focused on event-based sentiment analysis from tweets and discovered that the tweet events' sentiment influences financial events for the community/company.

Although much work has been carried out, social media sentiment analysis is still challenging due to the limited information in one single message of social media, in addition to indirect expressions, as well as the noisy information within social media such as advertisements and news. In all the of above prior art, the importance of noise filtering and tweet pre-processing is mentioned. Pre-processing includes a wide range of approaches to reduce noise.

All the research in the state of the art is focused on a subset of tweets out of the overall tweet stream, e.g. sentiment analysis via keywords search from tweets for relevant brands/products. When we consider the whole tweet stream to investigate the general (overall) population sentiment, we need to consider all tweet-carrying sentiment information. We need to filter out tweets with no sentiment bearing content, e.g. advertisement, news and spam. As mentioned in the state of the art, filtering is very important, yet challenging. Filtering noise from a sub-stream of tweets related to a particular subject matter, event or brand has been difficult; filtering noise from the whole general tweet stream is even harder. Hence, the first challenge is how to filter out the information with no sentiment element from the general tweet stream. After filtering all non-sentiment tweets, sentiment analysis for general tweets remains more challenging than sentiment analysis for a specific application, this is due to the vast diversity of vocabularies and expressions used across different areas. It gets more complex when the tweet does not belong to any single defined subject matter, e.g. "Good day today," and has sentiment information but not for any particular area. However, this information is indicative of the individual being happy and contributes to the overall happiness index for the whole population. On the other hand, given sentiment analysis from overall general tweets, obtaining sentiments for specific topics, brands, events, etc. is merely a tweet sub-selection task using either keywords or machine learning techniques to extract tweets from the overall tweet stream.

In summary, a sizeable amount of work has been done on sentiment analysis from social media, which we cannot list fully. However, as stated in the recent survey (Yue et al. 2018), most work focuses on consumer and population opinion mining for companies conducting surveys about corresponding products and services, national security or public views on events such as elections. Very little work is done on general populace sentiment analysis due to the challenges mentioned in Sect. 1. This paper proposes a general system for population happiness index monitoring using sentiment analysis from social media data (Twitter) which is introduced in Sect. 3.

The proposed system

There are five main components: tweet smart harvester, tweet filter, sentiment classifier, tweet sub-stream selector and visualiser (as shown in Fig. 1).

Smart harvester

We are using the Twitter REST APIs provided by Twitter for tweet harvesting. Java is used for the system development, and Twitter4j is used to connect to the official Twitter REST APIs. Twitter REST APIs provide 100 * 180 tweets per 15 min, which adds up to maximum 1,728,000 (over 1 million) tweets per day. However, we can only access tweets that have been indexed and made accessible to the public. Twitter states that not all tweets are indexed and only a subset of the whole twitter stream is provided when using Twitter REST APIs. Twitter also provides paid for subscriptions that can allow various additional levels of access. However, at a country/nation level a portion of the overall twitter stream is still huge considering the total amount of tweets within that country/nation. Hence, as long as the harvester is able to access and harvest a constant amount (percentage) of the total tweets, the tweets we harvested are representative of the overall happiness trend/index of the whole society. Experimentally, the harvester obtains 0.3 to 0.4 million tweets daily in UAE and about 50 thousand tweets daily within Abu Dhabi. The streams are very stable as shown in Fig. 2. All tweets acquired contain information about author, location, time and tweet content (text). The implementation used Twitter REST APIs for automatically requesting tweets and then pushing them into Elasticsearch from time to time. It must be noted that when harvesting tweets, Twitter terms and condition for use must be followed and measures need to be taken to ensure compliance.

Filter

This is a very important component of the proposed system, especially when dealing with general tweets and not a particular subset, such as theme, topic, brand or event. There are three stages of filtering as shown in Fig. 1.

Stage 1

First, tweets are filtered by author, auto-messages (automatically or semi-automatically generated tweets by a system or agency), mobile number, no valid information and auto-prayers (automatically generated prayer

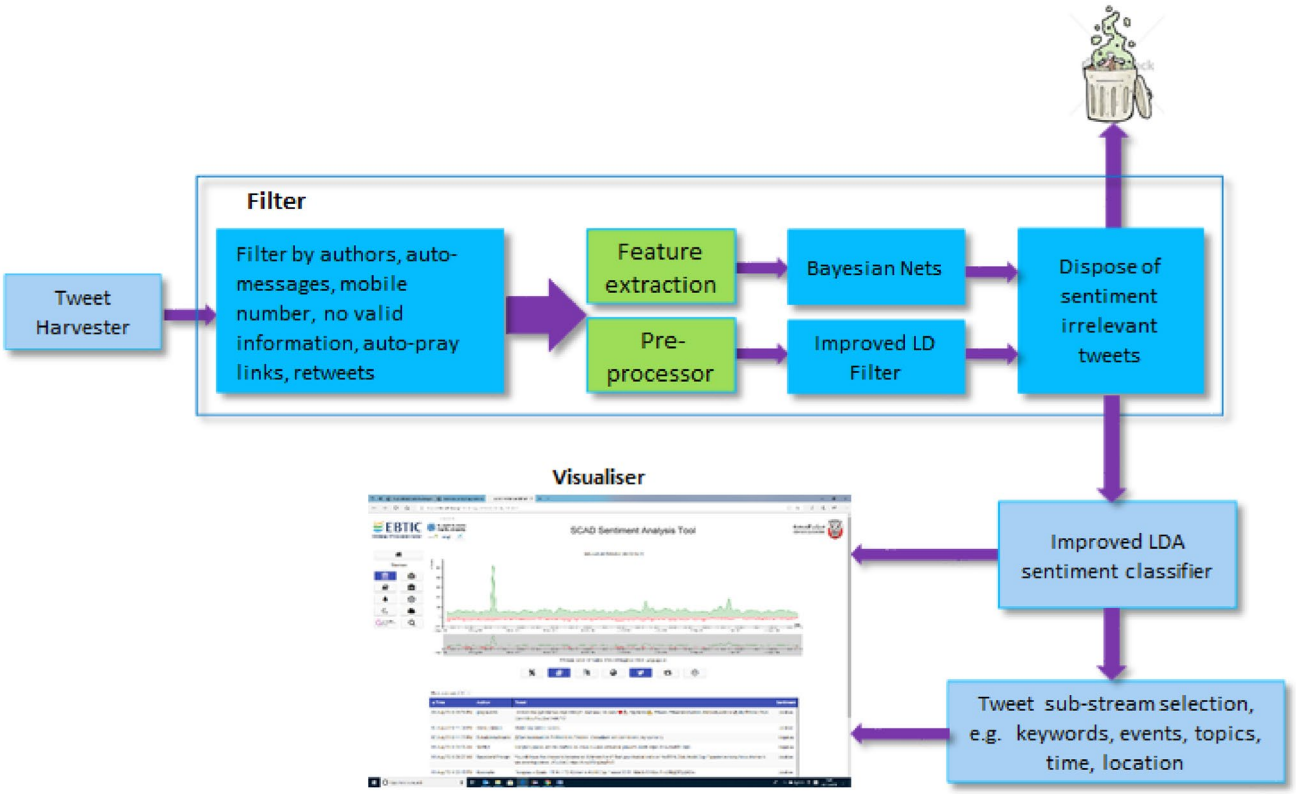

Fig. 1 Happiness index system from tweets

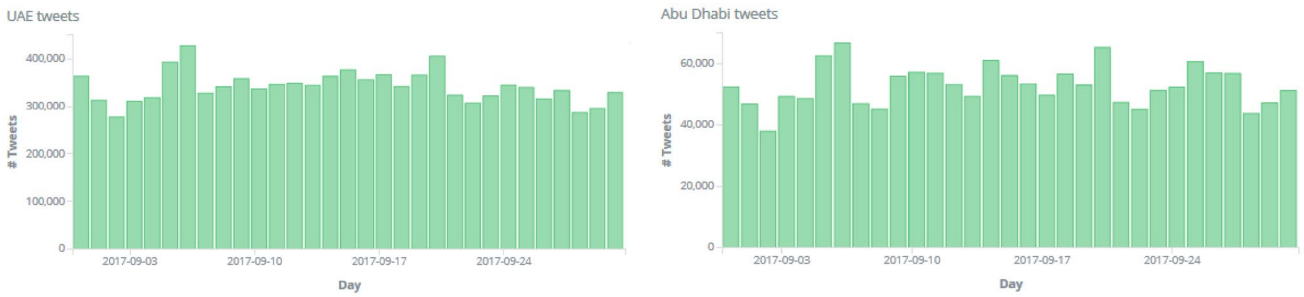

Fig. 2 Daily number of harvested tweets (for any language)

tweets). Company or media authors/accounts will not be considered for population sentiment, and they are manually tagged as non-personal accounts and hence do not contribute to population sentiment. To identify these non-personal accounts, a manual process is required. However, this is an incremental process, new non-personal accounts will always be added incrementally once identified, which means less and less manual involvement is needed through time. Auto-messages might come from personal accounts, but are machine generated and do not contribute to the general sentiment analysis either, e.g. "one user followed me, two users unfollowed me." In addition, we also observed that tweets with mobile numbers tend to be advertisements or conversations with customer services accounts, which we also consider as not carrying sentiment element and filter out. Tweets including no valid information (e.g. tweets only containing URLs or/and meaningless strings) are filtered out too. In the Arab world, many people use services to automatically generate tweets that have links to prayer websites. We consider these automatically generated prayer tweets, as not carrying sentiment information; hence, they will be also filtered out in the first filtering stage.

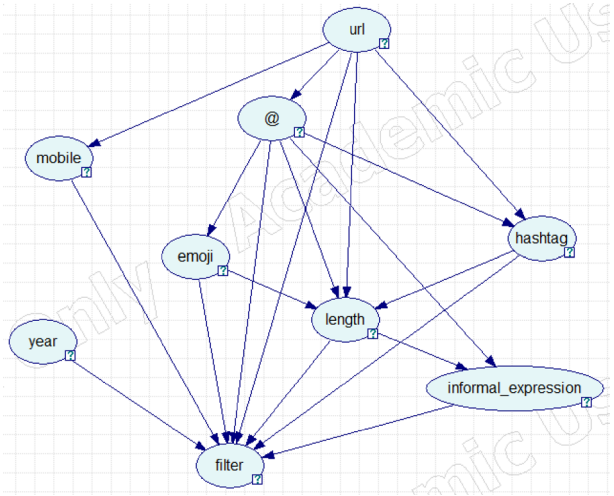

Fig. 3 Bayesian network tweet filter model

Stage 2

The remaining tweets will be input into Bayesian network (BN) and latent Dirichlet allocation (LDA) at the same time to be classified as sentiment tweets and non-sentiment tweets. The inputs to BN are the context properties of tweets, such as the number of included URLs/hyperlinks, mentions (@), hashtags (#) and emoji, as well as the valid length of the tweets (that is, the length after removing URLs and @). Furthermore, we input whether the tweets include telephone/mobile numbers, year information and informal expressions (more than 2 continuous repeated characters). Figure 3 shows the generated model of the BN tweet filter. The model (both structure and parameters) is trained/optimised from a batch of tagged tweets as training data. Hence, BN looks into the properties/context of a tweet, while LDA looks into the text/content of a tweet. By combining the results from BN and LDA, we consider both the tweet content/context and other characteristics of the tweets, e.g. author, URLs, etc. to identify whether a tweet carries sentiment element or needs to be filtered out. BN is good at dealing with structured data, i.e. characteristics of tweets, while LDA is good at dealing with text, i.e. the content of a tweet. Tweets classified as non-sentiment tweets by both BN and LDA will be discarded at this stage.

Stage 3

Now, the remaining tweets are then input into the improved LDA filter to classify sentiment-relevant tweets and sentiment-irrelevant tweets (we call junk) and the other to classify positive, negative, other (neutral or junk). Stemmers are proved to be an important part of pre-processing.

The number of tweets filtered out by stage 1 (filtered by authors, auto-messages, mobile numbers, no valid information and auto-prayers) is shown in Fig. 4. We can see a small percentage (less than 5%) are filtered out during stage 1 filtering. Adding more filtered author will increase this percentage, but we need to balance the manual effort and the filtering gained at this stage. Because stage 1 filtering is rule-based, all junk tweets from stage 1 are defined as junk; therefore, the accuracy in this stage is 100%.

The number of tweets filtered out by stage 2 is shown in Fig. 5. We can see a big chunk of tweets are filtered out. We manually checked a sample of filtered out tweets in stage 2 and demonstrated that almost all tweets classified as junk are indeed junk (99% accuracy), which means we lose very little information by using stage 2 filtering. We need to ensure that we do not lose too much sentiment information (positive and negative tweets) when filtering out junk tweets.

After the filters of stages 1 2, the most difficult task is sentiment classification (positive, neutral, negative and junk). We will cover the details and present the results in the experiment sections.

The filter obtains the latest tweets (which have not yet been tagged as junk/relevant) from Elasticsearch and the generated model calculates the probability of the tweets being junk and writes the results back to Elasticsearch.

LDA sentiment classifier

The core element of our happiness index monitoring system is tweet classification, which is used for both filtering (as mentioned in last paragraph) and sentiment analysis (to classify sentiment-relevant tweets into positive, negative, neutral

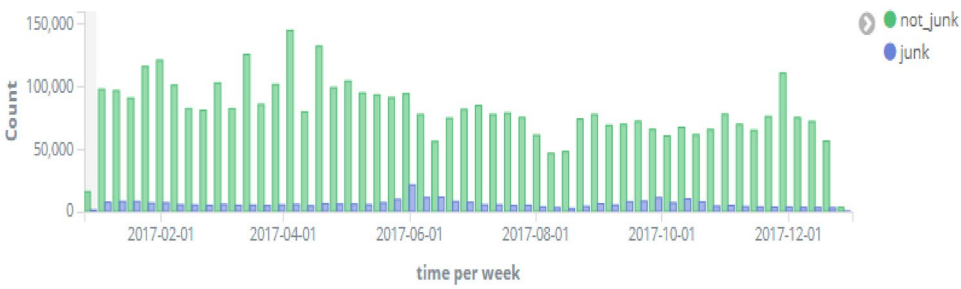

Fig. 4 Stage 1 filtering (English tweets in UAE)

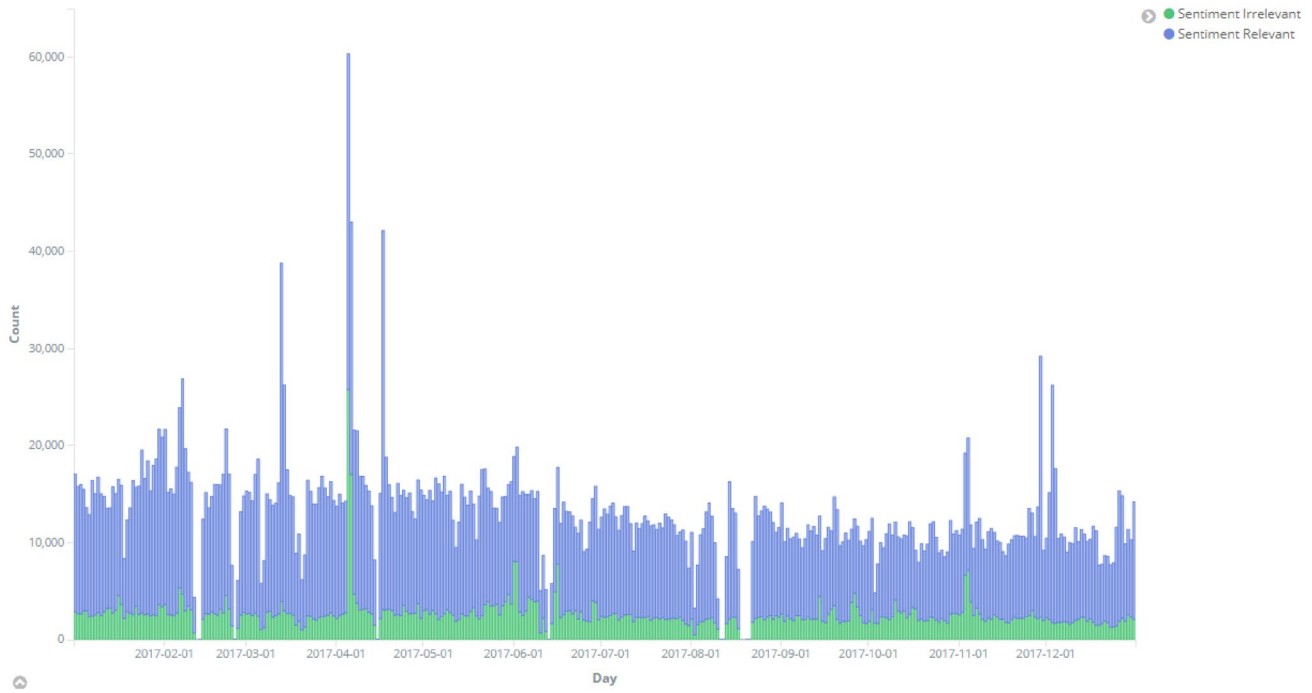

Fig. 5 Stage 2 filtering (English tweets in UAE)

and junk). This can be also considered as a combination of sentiment analysis and additional stage of filtering. However, this is a combined model, and we do not separate this as a separated filtering stage. The LDA (for filtering and sentiment classification) needs a pre-processing module. In the pre-processing, in addition to removing invalid information and text normalisation, a stemmer is also applied. Tweet classification accuracy is key. We applied our improved version of LDA for tweet classification which has achieved better accuracy for general classification problems for tweets (Wang et al. 2017; Wang and Al-Rubaie 2015).[2] The pseudocode of the filtering and sentiment analysis process is shown in Fig. 6. The detailed algorithm for our improved LDA is included in the references (Wang et al. 2017; Wang and Al-Rubaie 2015)[3] and pseudocode is included in reference (Wang and Al-Rubaie 2015). A patent application has been filed on the improved LDA method, and more details are in reference[4]. Our proposed improved LDA is also compared with existing algorithms applied to benchmark tweet sentiment datasets, and the results are discussed in Sect. 4.

Our proposed system continuously collects the latest tweets (which has not been tagged for sentiment) from Elasticsearch and calculates the probability of the sentiment (positive, negative or neutral) for these latest tweets and stores the results back in Elasticsearch.

Tweet sub-stream selector

In addition to providing sentiment from the overall tweet stream, sentiment for topics and events of interest can also be provided by narrowing the stream down by topic selection, hashtag selection, location and time period selection or keyword search. The tweet sub-stream selector provides a comprehensive capability to view different sentiment constituent from different areas/topics to compose the overall happiness index. In other words, our proposed system helps understand not only the overall national sentiment, but also the components contributing to the national sentiment and their effect, more precisely, the underlying reasons for the national sentiment. Hence, potentially providing an insight into the factors that influence the index, e.g. the population is generally happy but not that happy with public transport, so improving the public transport might help with improving the overall happiness index.

The sub-stream selector is built into the visualiser to select sub-streams of interest by searching from Elasticsearch.

[2] https://patents.google.com/patent/US20170293597A1/en.

[3] https://patents.google.com/patent/US20170293597A1/en.

[4] https://patents.google.com/patent/US20170293597A1/en.

```
Obtain all newly harvested tweets from Elasticsearch
for each tweet T

#stage 1 filtering
if the tweet is from one of the list of defined companies/organisations:
tag it as junk and continue for next tweet;
else if the tweet includes any auto-messages tag, e.g. thumb for youtube, auto-link to pray websites:
tag it as junk and continue for next tweet;
else
pre-process to remove uninformative strings/characters such as links, mentions, etc.
if no words or emoji left:
tag it as junk and continue for next tweet;
else
continue to stage 2 filtering

#stage 2 filtering
extract features from the tweet (whether include url, hashtag, mention, emoji, mobile number, etc. as shown in Figure 3)
input the extracted features into the bayes net model in Figure 3 and get the probability_filter (as junk)
if probability_filter > 0.5:
tag it as junk and continue to next tweet;
else:
continue with stage 3 filtering

#stage 3 filtering
input the pre-processed tweet into filter LDA model to classify the tweets as either sentiment relevant or sentiment irrelevant

#sentiment analysis
input the preprocessed tweet into sentiment LDA model to classify the tweets as positive, negative, neutral, or junk

#write the result back to Elasticsearch
push the result (positive, negative, neutral, or junk) back into Elasticsearch
```

Fig. 6 Pseudocode for tweet filtering and sentiment analysis

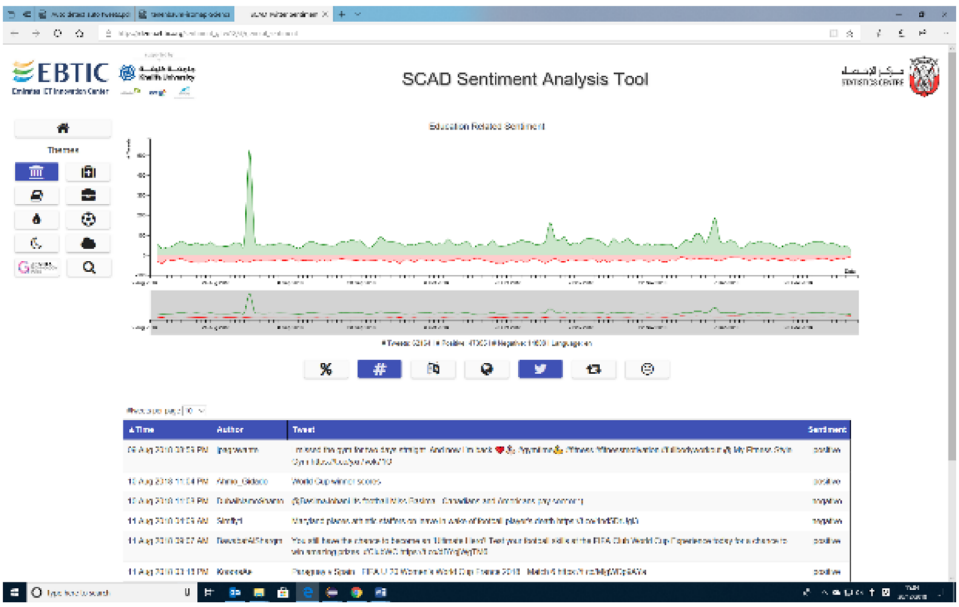

Fig. 7 General sentiment visualisation in Abu Dhabi

Table 1 Datasets used for comparison

Dataset	Total	Positive	Negative	Neutral
OMD[a]	1081	393	688	
HCR[a]	1354	397	957	
STS-Gold[a]	2034	632	1402	
STS-manual (Wang and Al-Rubaie 2015)/testdata.manual.2009.06.14 (Blei et al. 2003) also referred as STS test or STS-test in the reference)	498	182	177	
Twitter-sander-apple2 (Blei et al. 2003)	479	163	316	139
Twitter-sander-apple3 (Blei et al. 2003)	988	163	316	
Twitter dataset online (Blei et al. 2003)	2000	1000	1000	509
Sanders (Wang and Al-Rubaie 2015)	3237	478	481	2278

[a]https://patents.google.com/patent/US20170293597A1/en

The visualiser

The comprehensive visualiser shows tweet sentiments through time and across location, for both general and sub-stream sentiment through tweet sub-stream selection as shown in Fig. 7.

Improved LDA classifier for sentiment analysis for benchmark datasets

Tweet sentiment analysis is a difficult problem of text classification. Text classification accuracy is the key for a happiness index monitoring system and is used for both the filtering process (classify all tweets into sentiment-relevant and sentiment-irrelevant tweets) and sentiment analysis (classify sentiment-relevant tweets into positive, negative, neutral and junk).

From the state of the art, various machine learning techniques for text classification are applied to sentiment analysis including naïve Bayes, maximum entropy (MAXENT), support vector machine (SVM), artificial neural network (ANN) and latent Dirichlet allocation (LDA). LDA (Blei et al. 2003) was introduced by Blei and is a statistic model for classification based on word–topic frequency distribution. When LDA is applied directly to tweets, we lose the accuracy due to the shortness of tweets. We proposed a method, the improved LDA, to tackle this difficulty which is the subject of a pending patent application.[5] Compared with other machine learning methods, no keywords are needed to be pre-defined for LDA itself which makes it more attractive for text mining. We proved the higher accuracy of our proposed improved LDA over SVM and other algorithms in our previous work (Wang et al. 2017; Wang and Al-Rubaie 2015).

We are using our improved LDA for tweet sentiment analysis, i.e. classification of positive, negative, neutral and junk, and as filter, i.e. classification of sentiment-relevant and sentiment-irrelevant tweets. Our proposed improved LDA is also compared to state-of-the-art algorithms for benchmark tweet datasets.

To prove the advantages of using our improved LDA for sentiment analysis in comparison with the state-of-the-art algorithms, we applied our improved LDA to a set of benchmark datasets used in the literature as shown in Table 1.

We first apply our proposed method to benchmark database OMD (Wu et al. 2016), STS-Gold (Wu et al. 2016) and HRC database (Ren et al. 2016; Wu et al. 2016). The comparison of our proposed method with existing methods and baseline methods are shown in Table 2 and Figs. 8, 9 and 10. When comparing with the results in (Wu et al. 2016), our proposed improved LDA outperforms all their algorithms for all datasets, mainly because their methods are lexicon-based, while our proposed improved LDA is machine learning based. When compared with the best result of other machine learning techniques with ensemble learning (Troussas et al. 2016), our proposed improved LDA is comparative to their accuracy (0.8579 using ours and 0.8774 using ensemble-based methods for OMD dataset, 0.8629 using ours and 0.8510 using ensemble-based methods for HCR dataset and 0.9130 using ours and 0.8902 using ensemble-based methods for STS-Gold dataset). Their ensemble learning is based on the combined/best results of popular machine learning methods (NB, SVM, KNN and C4.5) which proves that ensemble learning increases the accuracy of machine learning methods in general. Therefore, we expect the accuracy for our proposed improved LDA to increase when embedded in ensemble learning techniques, which is part of our future work. Our proposed improved LDA is also better than Zhao's method (Pandey et al. 2017) for STS-Gold dataset. (Result for other datasets is not available in their paper for comparison here.)

[5] https://patents.google.com/patent/US20170293597A1/en.

Table 2 Accuracy comparison for OMD and HRC datasets

	Dataset					
	OMD (Wu et al. 2016)		HCR (Wu et al. 2016)		STS-Gold (Wu et al. 2016)	
	Accuracy	*F*-measure	Accuracy	*F*-measure	Accuracy	*F*-measure
Baseline-ML (Wu et al. 2016)	0.5162	0.516	0.4742	0.4741	0.5747	0.5746
Baseline-SL (Wu et al. 2016)	0.5282	0.5269	0.4882	0.4858	0.5664	0.5592
Baseline-SS (Wu et al. 2016)	0.6679	0.614	0.6699	0.596	0.8132	0.7856
SentiWordNet-Mdn (Wu et al. 2016)	0.6161	0.6061	0.6617	0.5364	0.6937	0.6583
SentiWordNet-Pvt (Wu et al. 2016)	0.6531	0.6246	0.6329	0.5451	0.707	0.658
SentiWordNet-Hbd (Wu et al. 2016)	0.6698	0.6578	0.6322	0.5453	0.7114	0.6689
MPAQ-Mdn (Wu et al. 2016)	0.6698	0.5936	0.7068	0.5889	0.762	0.7148
MPAQ-Pvt (Wu et al. 2016)	0.704	0.6562	0.6846	0.5539	0.7606	0.7129
MPAQ-Hbd (Wu et al. 2016)	0.7058	0.6694	0.6832	0.5635	0.7566	0.7221
Thewall-Lexicon-Mdn[a]	0.6707	0.5975	0.6869	0.5428	0.7974	0.7615
Thewall-Lexicon-Pvt (Wu et al. 2016)	0.6901	0.6409	0.6706	0.5335	0.7984	0.7586
Thewall-Lexicon-Hbd (Wu et al. 2016)	0.6984	0.662	0.6699	0.5422	0.8033	0.7752
Our improved LDA	0.8579	0.8585	0.8629	0.86	0.9130	0.9129
Christos's ensemble-based methods (Troussas et al. 2016)	0.8774	NA	0.8510	NA	0.8902	NA
Best Result in (Nguyen and Jung 2017)	0.811	0.803	0.749	0.721	0.818	0.806
Zhao's method (Pandey et al. 2017)	NA	NA	NA	NA	0.8007	NA

[a]https://patents.google.com/patent/US20170293597A1/en

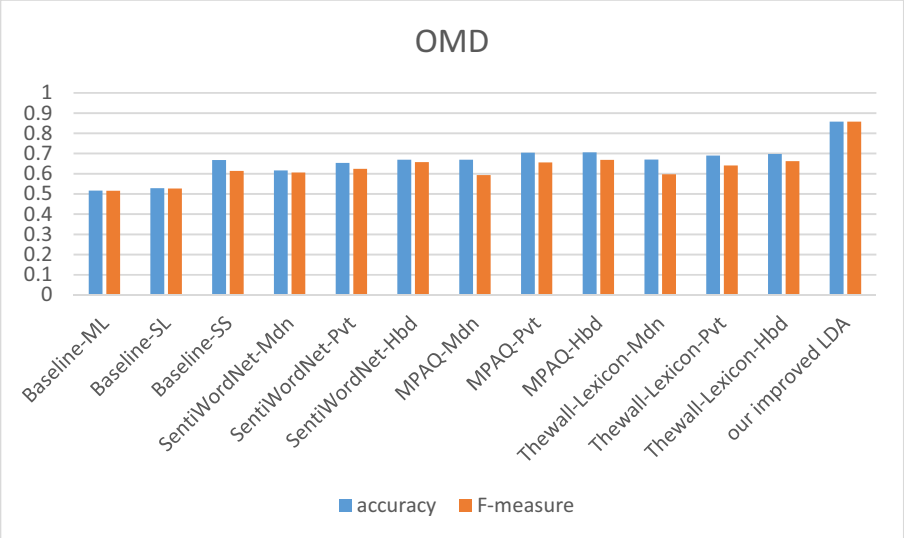

Fig. 8 Accuracy comparison for OMD dataset

Then we apply our proposed improved LDA to STS and Sanders datasets (Ren et al. 2016) for accuracy comparisons with different systems. We apply the same strategy of fivefold cross-validation to compare our accuracy with the result in Ren et al. (2016) and Pandey et al. (2017), as shown in Table 3 and Fig. 11. We can see our proposed improved LDA obtains the best accuracy, which is overall 2% better in accuracy for both datasets of STS and Sanders. When compared with Zhao's method (Pandey et al. 2017), our proposed method obtains better accuracy for STS-manual dataset, and the accuracy for Sanders dataset is not available from Zhao's paper (Pandey et al. 2017).

Other tweet datasets we use for accuracy comparisons are: test data manual 2009.06.14 (Saifa et al. 2016) (also referred as STS-manual test (Ren et al. 2016)), Twitter-sander-apple2 (Saifa et al. 2016), Twitter-sander-apple3 (Saifa et al. 2016) and Twitter dataset online (Saifa et al. 2016). The accuracy results are shown in Table 4 and Fig. 10. From Table 4, we can see that our proposed improved LDA achieves the best accuracy for the datasets of test data manual 2009.06.14

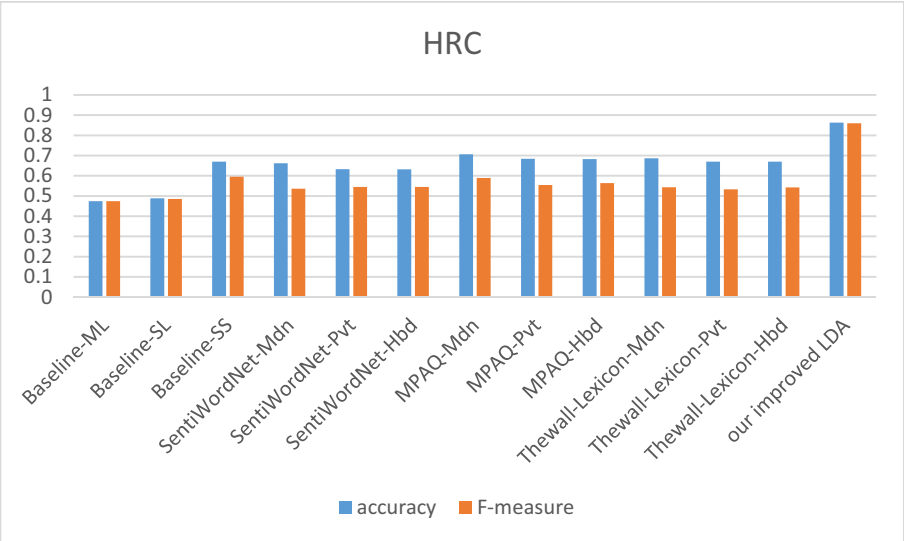

Fig. 9 Accuracy comparison for HRC dataset

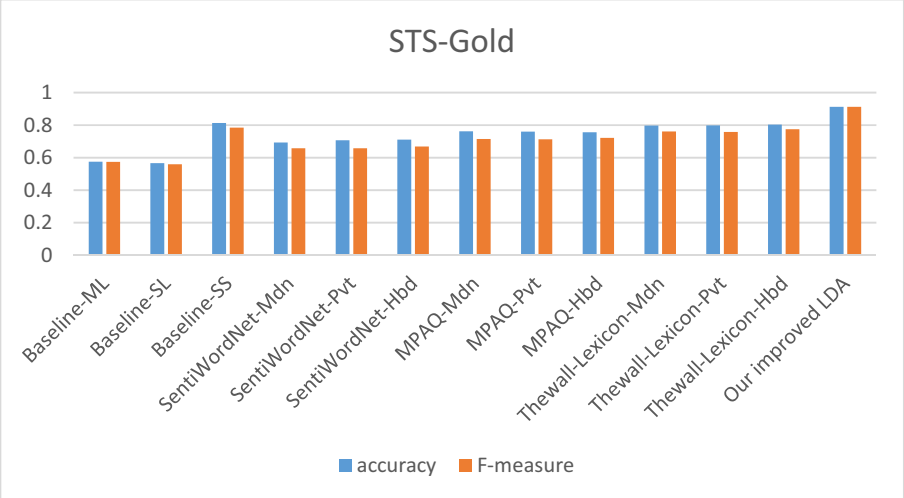

Fig. 10 Accuracy comparison for STS-Gold dataset

Table 3 Accuracy comparison for STS and Sanders datasets

Methods	STS-manual (Wang and Al-Rubaie 2015)	Sanders (Wang and Al-Rubaie 2015)
LS (Wang and Al-Rubaie 2015)	0.8228	0.8338
Log (Wang and Al-Rubaie 2015)	0.8218	0.8366
SVM (Wang and Al-Rubaie 2015)	0.7969	0.8218
NB (Wang and Al-Rubaie 2015)	0.8375	0.8207
DistSup (Wang and Al-Rubaie 2015)	0.7673	0.7297
ESSA (Wang and Al-Rubaie 2015)	0.7536	0.7421
ESLAM (Wang and Al-Rubaie 2015)	0.8698	0.8275
NRC-Canada (Wang and Al-Rubaie 2015)	0.8537	0.8406
Coooolll (Wang and Al-Rubaie 2015)	0.8543	0.8396
HSK-LS (Wang and Al-Rubaie 2015)	0.8729	0.8628
HSK-Log (Wang and Al-Rubaie 2015)	0.8817	0.8636
HSK-SVM (Wang and Al-Rubaie 2015)	0.8838	0.8606
our improve LDA	0.9032	0.882
Zhao's method (Jianqiang and Xiaolin 2017)	0.8161	NA

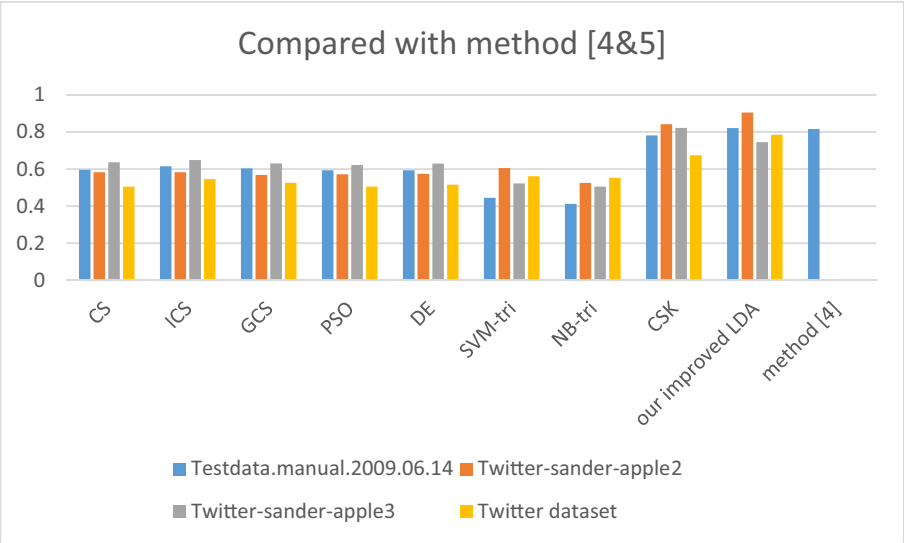

Fig. 11 Accuracy comparison for OMD and HRC datasets

Table 4 Accuracy comparison for OMD and HRC datasets

Methods	testdata.manual.2009.06.14	Twitter-sander-apple2	Twitter-sander-apple3	Twitter dataset
CS (Blei et al. 2003)	0.5954	0.5828	0.6362	0.5058
ICS (Blei et al. 2003)	0.6148	0.5829	0.6485	0.5463
GCS (Blei et al. 2003)	0.6041	0.5681	0.6307	0.526
PSO (Blei et al. 2003)	0.5928	0.5724	0.6217	0.5055
DE (Blei et al. 2003)	0.5936	0.5745	0.6301	0.516
SVM-tri (Blei et al. 2003)	0.4447	0.6051	0.5227	0.5615
NB-tri (Blei et al. 2003)	0.4123	0.525	0.5053	0.5525
CSK (Blei et al. 2003)	0.7817	0.8416	**0.8221**	0.6745
Our improved LDA	**0.9032**	**0.9048**	0.7836	**0.7845**
Method (Krouska et al. 2016)	0.8161	NA	NA	NA

Bold values indicate best result for the corresponding column across different methods/algorithms

Table 5 Accuracy comparison for gold stanford dataset

Method (Singh and Kumari 2016)	Precision			Recall			*F*-score		
	Positive	Negative	Other	Positive	Negative	Other	Positive	Negative	Other
Naïve Bayes	0.267	0.352	0.831	0.598	0.648	0.449	0.369	0.456	0.583
Bayes Net	0.38	0.348	0.81	0.295	0.639	0.678	0.332	0.45	0.636
DMNB	0.679	0.739	0.805	0.381	0.52	0.935	0.488	**0.611**	0.865
SMO	0.543	0.614	0.82	0.475	0.537	0.866	0.507	0.573	0.842
Hyperpipes	0.415	0.544	0.755	0.308	0.289	0.88	0.353	0.377	0.812
Random forest	**0.857**	**0.881**	0.718	0.088	0.167	**0.989**	0.159	0.276	0.832
Our improved LDA	0.453	0.496	**0.945**	**0.64**	**0.74**	0.89	**0.532**	**0.594**	**0.89**

Bold values indicate best result for the corresponding column across different methods/algorithms

(Ren et al. 2016; Saifa et al. 2016), Twitter-sander-apple2 (Saifa et al. 2016) and Twitter dataset. When the proposed improved LDA is applied to Twitter-sander-apple3 (Saifa et al. 2016), we achieve an accuracy of 0.7836 which is the second best among all compared methods. Twitter-sander-apple3 has 3 categories instead of 2 as in the other datasets. This is expected, as when using training-based classifiers the accuracy decreases when the number of categories increases.

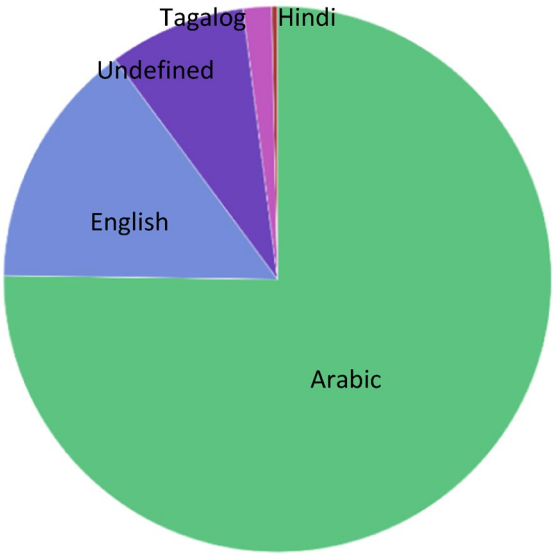

Fig. 12 Percentage of tweets per language in Abu Dhabi

Zimbra et al. (2016) tested the performance of various algorithms for sentiment analysis by using the Sanders Twitter dataset. The accuracy comparison in precision, recall and *F*-score is shown in Table 5. Our proposed improved LDA and random forest tend to achieve better results than other methods. Random forest obtains the best precision but very low recall for positive and negative categories, which makes random forest less useful here because it fails to identify most actual positive and negative tweets. The *F*-score for positive and negative categories from random forest is very low due to their low recall rate. Our proposed improved LDA achieves the best *F*-score for positive and other categories, and the second best accuracy for the negative category.

General sentiment analysis

Accuracy for general sentiment analysis

We are running an online harvester in the background to continuously harvest tweets from Abu Dhabi and UAE. These tweets are pre-filtered using the rules of authors, auto-messages, no valid information and auto-prayers (as Stage 1 of filtering). The tweets are then filtered by Bayes Nets (as Stage 2 of filtering). Then, the proposed improved LDA is applied to tweets that are not pre-filtered as junk/sentiment-irrelevant. The proposed improved LDA is a supervised learning method which needs a training file to generate the model (Wang et al. 2017; Wang and Al-Rubaie 2015).[6] The training file contains a list of tweets (after filtering) that are manually tagged as positive, negative, neutral or junk/irrelevant, because the remaining tweets still contain some junk tweets that could not be filtered out in Stages 1 and 2 filtering. Sentiment analysis is a difficult task for classification, and it is even more difficult for general sentiment analysis from all tweets. The difficulty stems from the ambiguity of sentiment expressions (which are even more ambiguous in tweets due to their informal nature) and subjective interpretation from one individual to another.

To generate the training and testing data, two types of training schemes are involved: individual tagging and group tagging. In group tagging, each of the taggers are requested to tag all the tweets in the set, so each tweet is tagged by every tagger. In individual tagging, each tweet is tagged only by one tagger. Group tagging is used to identify the consistency of tagging from different taggers, while individual tagging is used to extend the size of training data.

Twitter identifies and tags tweets with the language used, but in many Arab countries, the majority of tweets are either Arabic or English, e.g. percentage of tweets in Abu Dhabi as shown in Fig. 12. From Fig. 12, we can see that the combination of English and Arabic tweets covers about 90% of total tweets in Abu Dhabi UAE.

However, some Arabic tweets include English words occasionally as well. If the majority of words in one tweet are Arabic words, then twitter tags it as Arabic. Our proposed improved LDA is language independent and hence is able to deal with both Arabic and English. We generate separate models for Arabic and English for better accuracy using the same technique (our improved LDA) because English and Arabic tweets use different vocabulary sets. English words that frequently occur in Arabic tweets can be caught and built in the generated Arabic model as well. In theory, a unique model to deal with both Arabic and English is possible but might lose some accuracy due to more variation in the total vocabulary when combining Arabic and English vocabularies.

For English tweets, we train our improved LDA model by using 12 k tagged tweets (Dataset A, from individual training) and test it using a different set of 300 random tweets with group tagging (Dataset B, all these 300 tweets are tagged by 4 taggers and a voting scheme among these four taggers is used to decide the final tagging for testing). The 4 taggers are native English speakers from different English-speaking countries to cover general English usages.

To highlight the ambiguity of sentiment expressions in tweets and the difficulty of sentiment analysis from tweets, we test the consistency from our 4 taggers for the random 300 tweets (Dataset B) and the consistency percentage from all 4 taggers is only 46%. The accuracy from our LDA is between 44 and 52% when changing the training data size (600 to 12 k). Therefore, our proposed improved LDA

[6] https://patents.google.com/patent/US20170293597A1/en.

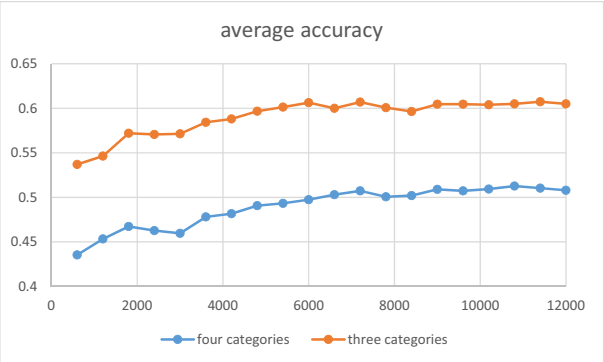

Fig. 13 Accuracy increases with the size of training data for general sentiment analysis

Table 6 Detailed results for LDA sentiment classification for English

	Positive	Negative	Neutral	Junk
Positive	103	12	37	7
Negative	12	42	44	3
Neutral	24	23	75	13
Junk	8	2	26	34

obtains comparable or better accuracy when compared to manual tagging. Most of the disagreements between taggers are related to neutral v. positive and neutral v. negative, which indicates that tagging itself depends on personality: positive thinking or negative thinking. This double confirms the challenges of sentiment analysis. In addition, there is also some confusion between positive and negative in manual tagging, e.g. "I miss you" is tagged as positive by one tagger and negative by another, and both are explainable. This further shows the challenges of general sentiment analysis in comparison with topic/product/event/service specific sentiment analysis.

We now investigate how the size of training data influences the accuracy. Among the pool of randomly sampled 12 k tagged tweets (Dataset A), we sample the tagged tweets from 5% (600) tweets to 100% (12 k) tweets as the training sets to generate the model. We tested the generated model by using Dataset B, a different set of 300 tweets tagged by 4 taggers and a voting scheme. We run the experiment 10 times, and the average accuracy according to the size of training data is shown in Fig. 13. The blue line shows the accuracy of four categories (positive, negative, junk and neutral). Most of the time we do not need to distinguish junk from neutral, because neither contribute to sentiment change. The orange line shows the accuracy of three categories (positive, negative and [junk or neutral]). We can observe the trend of accuracy increasing along the increase of the training dataset, especially at the beginning of training

Table 7 Detailed results for LDA sentiment classification for English

	Positive	Negative	Neutral	Junk
Positive	260	40	29	10
Negative	80	54	19	6
Neutral	50	28	34	21
Junk	13	5	12	25

data size increase. The accuracy stays stable after 6 k of training data.

In the real world, for population sentiment analysis, neither neutral nor junk contribute to sentiment analysis. Therefore, we can combine junk and neutral as irrelevant tweets to be filtered out, which leads to 3 categories for the classifier (positive, negative and irrelevant as a combination of junk and neutral). We can see the accuracy is around 60% with three categories, which again shows the difficulty of sentiment classification from general tweets with no constrictions.

Table 6 shows one set of test results of 300 random tweets. The green-coloured figures are those correctly classified; the blue-coloured figures are those wrongly classified but can be considered to compensate for each other to make the overall statistics correct; the grey-coloured figures are those wrongly classified between junk and neutral which do not change the statistics for population sentiment changes; hence, they can be ignored. The most impacting errors in classification that affect the performance and use of the system are the incorrect classification between neutral and positive, and between neutral and negative. Further research is needed to address this area. We can see the accuracy from Table 6 is 54.6% for four categories' classification: positive, negative, neutral and junk, and 63.0% for three categories' classification: positive, negative and other (neutral or junk).

For Arabic tweets, we use our proposed improved LDA for the Arabic training data (the same process as for English stated above), and the accuracy is shown in Table 7. The accuracy for four categories' classification: positive, negative, neutral and junk is 54.4%, and that for three categories' classification: positive, negative and other (neutral or junk) is 59.2%. We can see that our proposed improved LDA obtains similar accuracy for Arabic as it did for English. The slight loss of accuracy is caused by the complexity of the Arabic language itself. How to improve the pre-processing for Arabic to improve its classification accuracy further is part of our future work.

Accuracy for general sentiment analysis with emoji

People are using a lot of emoji in today's social media to express emotions. Emoji are a very valuable source of information for sentiment analysis, and they are intended

Table 8 Emoji frequencies for different sentiment

Emoji	Replaced text	Positive	Negative	Neutral	Others	Total
❤	Heavy_Black_Heart	156	8	42	37	243
✨	Sparkles	18	1	13	11	43
☺	White_Smiling_Face	16	5	8	10	39
☹	White_Frowning_Face	12	17	10	6	45
♥	Black_Heart_Suit	8	0	7	4	19
♨	Hot_Beverage	6	0	6	2	14
✋	Raised_Hand	5	0	5	3	13
✌	Victory_Hand	4	2	5	4	15
♻	Black_Universal_Recycling_Symbol	3	1	5	0	9
✈	Airplane	3	0	3	1	7
✔	Heavy_Check_Mark	2	0	2	3	7
❁	Snowflake	2	0	2	2	6
⛽	Fuel_Pump	2	0	3	1	6
☑	White_Heavy_Check_Mark	2	0	2	1	5
☁	Cloud	1	0	4	1	6

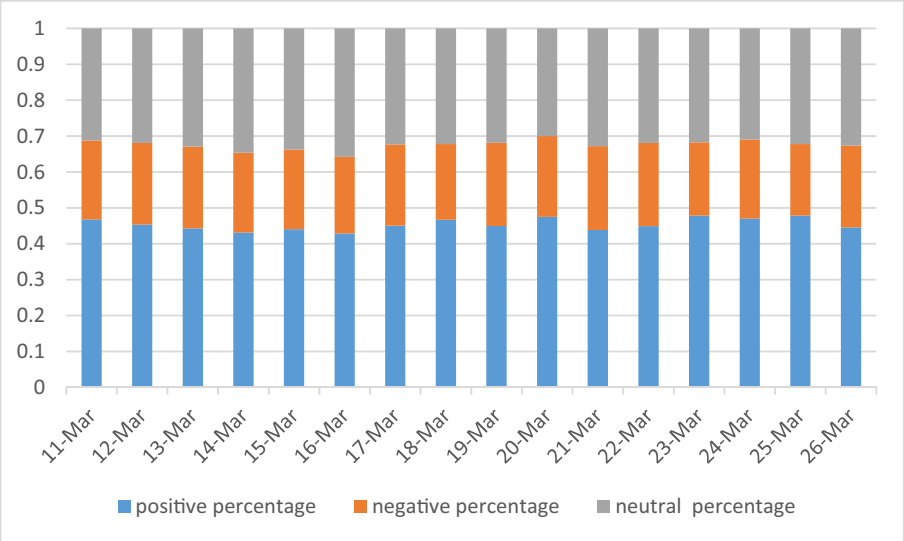

Fig. 14 Sentiment changes through 2 weeks in Abu Dhabi

to provide more recognisable representation of sentiment. Including emoji in sentiment analysis will greatly improve the accuracy. When we test the consistency of tagging from 4 taggers for another random 300 tweets (Dataset C), all of which include emoji, the consistency for the agreement for all 4 taggers increases to 52% (from 44% for general sampling from all tweets). This shows that emoji help reduce the sentiment vagueness.

However, there are no clear definitions for the sentiment of emoji, e.g. a heavy black heart can express a positive sentiment most of the time, but it occasionally also be negative. The sentiment represented by the emoji is also highly context (the message text used) dependant, which means that a lexicon of emoji for sentiment analysis will not work well. Table 8 shows the frequencies of different emoji occurring in 300 randomly sampled tweets with emoji. In our proposed method, as part of the pre-process, all emoji are replaced with meaningful semantic strings, which are defined by https://emojipedia.org/, e.g. ✨ is replaced with "Sparkles." After these replacements, all emoji are interpreted as machine understandable and meaningful strings. It helps the machine to better understand the sentiment when using emoji together with the words used in the tweets.

Our proposed improved LDA performs better when dealing with tweets containing emoji. When applying it to the same above mentioned 300 tweets with emoji (Dataset C), we obtain an accuracy of 68% compared to the consistency among four taggers of 52%. This shows that our proposed improved LDA obtains much better accuracy than the manual tagging when dealing with tweets containing emoji.

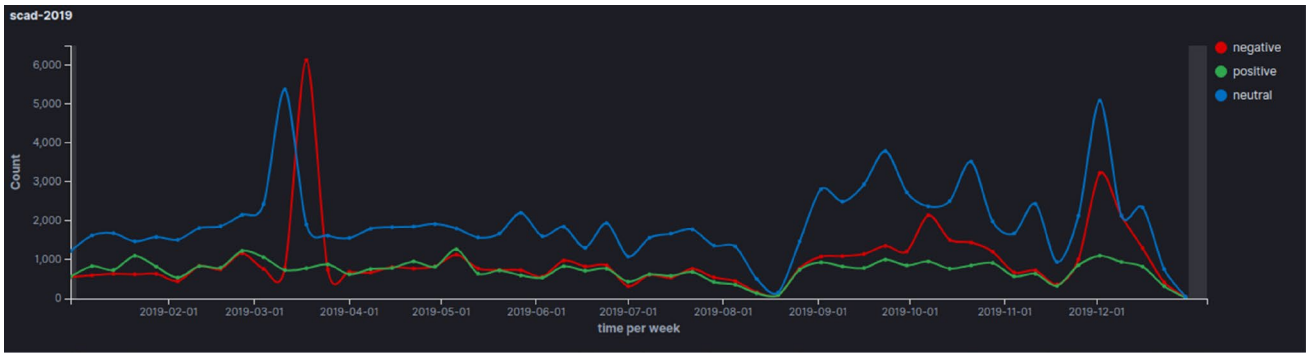

Fig. 15 Sentiment changes in 2019 in education in Abu Dhabi

Emoji together with text results in more assured sentiment judgement.

Sentiment monitoring in real time

We are using the results of our system for statistical analysis of populace sentiment changes over time. Figure 14 shows the sentiment changes in Abu Dhabi, UAE over two weeks, Saturday, 11 Mar 2017, to Sunday, 26 Mar 2017. Overall, people tend to be happier during weekends, 11 Mar (Saturday), 18 Mar (Saturday), 24 (Friday which is weekend in the UAE) Mar 2017 than weekdays, which is no surprise. In addition, Monday, 20 Mar 2017(in the blue box) is the international happiness day, we can see the peak of positive sentiment around that day as well. Another interesting finding is that the negative index stays stable through time and the sentiment changes tend to be a move from neutral to positive during weekends or event/festival periods. This can be attributed to the fact that certain people tend to be negative over prolonged periods of time due to personal, environmental and psychological circumstances; moreover, there are individuals who are chronically unhappy. Hence, we see this relatively steady percentage of negative sentiment. On the other hand, the majority of people tend to be more inclined towards being positive and happy saving specific factors, such as day of week, events and environmental factors. This latter portion of the populace makes up those who are showing either positive or neutral sentiments. While we acknowledge that there is a statistical change over time, seeing people move from one group to another, we can see a stable distribution throughout the results' period. This better explains the periodical populace sentiment changes and their correlation with days of the week (working days or weekends).

In addition, the proposed system provides an insight into the underlying reasons driving sentiment change in the populace through time, especially within a particular theme/area of interest, e.g. sentiment on the economy or education. For example, when we search for education-relevant tweets during 2019, we can see a peak of negative education-relevant tweets around March 20, 2019 (as shown in Fig. 15.). When we look at the content of these tweets, we find out that this is school exams period in the UAE and students are concerned or complaining about the exams. We can see a similar pattern during late November as well. Table 9 shows examples of tweets related to exams on 24 March as examples.

Conclusion and future work

In this paper, we proposed a system for general populace sentiment monitoring from social media. We highlighted and discussed the challenges of sentiment analysis from general tweets rather than specific streams such as brands, events, related tweets, as well as the challenges for sentiment analysis in Arab countries.

For general sentiment analysis, we need to accurately identify tweets with sentiment among all tweets. To do this we, proposed and used multi-level comprehensive filters that are able to filter out non-sentiment tweets accurately. An accurate filter is a prerequisite for accurate sentiment analysis.

A short message classifier is key for accurate filtering and accurate sentiment analysis. We used our proposed improved LDA for tweets (Wang et al. 2017; Wang and Al-Rubaie 2015; Blei et al. 2003)[7] sentiment analysis which is able to achieve the best accuracy for benchmark tweet datasets when compared with existing methods as shown in Sect. 4. For general sentiment analysis from all tweets, the proposed improved LDA is able to achieve similar or better accuracy for sentiment analysis than human tagging as shown in Sect. 5. A comprehensive tweet visualiser was also developed to show the users' sentiment changes, either in general or for particular interest(s), in geo-location distributions and through time.

[7] https://patents.google.com/patent/US20170293597A1/en.

Table 9 Example of tweets on education as negative

Time	Text	Sentiment
Mar 24, 2019 @ 23:55:34.000	it's not that i'm not motivated to study, i am, i really want to study but it's just my brain being so hyper and unable to focus	Negative
Mar 24, 2019 @ 23:54:56.000	idk why my brain doesn't want to study during the exams but during vacations my brain suddenly goes "whoa hey let's go study"	Negative
Mar 24, 2019 @ 23:54:52.000	Physics ❤❤❤❤❤❤❤❤❤❤	Positive
Mar 24, 2019 @ 23:54:01.000	i had a mini mental breakdown about math and physics but it's all good now	Negative
Mar 24, 2019 @ 23:11:07.000	RT @hend_mana: Pursuing a career in diplomacy has given me many opportunities to learn & interact w/ different cultures and ppl. If you're…	Neutral
Mar 24, 2019 @ 23:10:25.000	Allah yrz8ni a life without homework and exams	Negative
Mar 24, 2019 @ 23:09:57.000	RT @7amdaalman9oori: Thats it 5ala9 ill fail w im ok with that	Neutral
Mar 24, 2019 @ 21:59:49.000	Dad, can I go to my friend's house to study?	
me when I reach there: https ://t.co/xKrgxPve8E	negative	
Mar 24, 2019 @ 21:56:10.000	My heart goes out to Jelai. Let it out. You're wounded, and you may feel like you'll never recover. But it will pass. Keep breathing. Workout it out! It will give you extra endorphins. You go girl!	Negative
Mar 24, 2019 @ 21:52:49.000	My heart goes out to @jelaiandres_ Let it out. You're wounded, and you may feel like you'll never recover. But it will pass. Keep breathing. Workout it out! It will give you extra edorphines. You go girl!	Negative
Mar 24, 2019 @ 21:43:50.000	Thats it 5ala9 ill fail w im ok with that	Positive
Mar 24, 2019 @ 21:22:02.000	is this ethics course for real?	Neutral
Mar 24, 2019 @ 21:17:39.000	RT @ItsAmeeer: i'm actually convinced i cant pass an exam unless i cheat	Negative
Mar 24, 2019 @ 20:52:31.000	UM AMMAR INTERNATIONAL SCHOOL elliiiii f mushrif!!! Is fucking better than our school and the girls school:))!!	Negative
Mar 24, 2019 @ 20:51:53.000	My A level exams start in less than 2 months this is actually happening what the fuck https://t.co/S5MKY1ruoI	Negative
Mar 24, 2019 @ 20:43:46.000	RT @_tqubaisi: fuck physics	Negative
Mar 24, 2019 @ 19:59:16.000	RT @angelicnora: im failing chemistry this year:)	Negative

Although our improved LDA shows better accuracy in benchmark datasets, the accuracy for general sentiment analysis using real-world content is 68% and still needs to be improved. Deep learning and BERT (Devlin et al. 2019) are becoming popular for NLP tasks including text classification. However, BERT started for English and has been extended to other languages such as Arabic, i.e. AraBERT through Hugging Face (Wolf et al. 2020) very recently. Our future work will be investigating AraBERT for general sentiment analysis and how to combine it with our proposed method for better performance in accuracy.

References

Abdelwahab O, Bahgat M, Lowrance CJ, Elmaghraby A (2015) Effect of training set size on SVM and Naïve Bayes for Twitter sentiment analysis. In: 2015 IEEE international symposium on signal processing and information technology (ISSPIT), pp 46–51, 2015.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(2003):993–1022

Bravo-Marquez F, Frank E, Pfahringer B (2016) From opinion lexicons to sentiment classification of tweets and vice versa: a transfer learning approach. In: 2016 IEEE/WIC/ACM international conference on web intelligence, pp 145–152

Chiassi M, Skinner J, Zimbra D (2013) Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst Appl 40:6266–6282

Daniel M, Neves RF, Horta N (2017) Company event popularity for financial markets using Twitter and sentiment analysis. Expert Syst Appl 71:111–124

Deshwal A, Sharma SK (2016) Twitter sentiment analysis using various classification algorithms. In: 2016 5th international conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions), Sept. 7–9, 2016, AIIT, Noida, India, pp 251–257

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1, 2019

Furini M, Montangero M (2016) TSentiment: on Gamifying Twitter sentiment analysis. In: ISCC 2016 Workshop: DENVECT, 2016.

Gupta VS, Kohli S (2016) Twitter sentiment analysis in healthcare using Hadoop and R. In: 2016 international conference on computing for sustainable global development (INDIACom), pp 3766–3772, 2016.

Hassan A, Abbasi A, Zeng D (2013) Twitter sentiment analysis: a bootstrap ensemble framework, pp 357–364. SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013

Kanakaraj M, Guddeti RMR (2015) Performance analysis of ensemble methods on Twitter sentiment analysis using NLP technique. In: Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015), pp 169–170, 2015

Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N (2013) Ontology-based sentiment analysis of twitter posets. Expert Syst Appl 40:4065–4074

Krouska A, Troussas C, Virvou M (2017) Comparative evaluation of algorithms for sentiment analysis over social networking services. J Univ Comput Sci 23(8):755–768

Krouska A, Troussas C, Virvou M (2016) The effect of preprocessing techniques on Twitter sentiment analysis. In: 2016 7th international conference on information, intelligence, systems & applications (IISA), 2016

Nakov P, Rosenthal S, Kiritchenko S, Mohammad SM, Kozareva Z, Ritter A, Stoyanov V, Zhu X (2016) Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. Lang Resour Eval 50:35–65

Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V (2016) SemEval-2016Task4: sentiment analysis in Twitter. In: Proceedings of SemEval-2016, pp 1–18, 2016

Nguyen HL, Jung JE (2017) Statistical approach for figurative sentiment analysis on social networking services: a case study on Twitter. Multimed Tools Appl 76:8901–8914

Pandey AC, Rajpoot DS, Saraswat M (2017) Twitter sentiment analysis using hybrid cuckoo search method. Inf Process Manag 53:764–779

Peng Y, Moh M, Moh T-S (2016) Efficient adverse drug event extraction using Twitter sentiment analysis. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 1011–1018

Philander K, Zhong YY (2016) Twitter sentiment analysis: capturing sentiment form integrated resort tweets. Int J Hosp Manag 55:16–24

Porshnev A, Redkin I, Shevchenko A (2013) Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. In: 2013 IEEE 13th international conference on data mining workshops, pp 440–444

Qaisi LM, Aljarah I (2016) A Twitter sentiment analysis for cloud providers: a case study of Azure vs. A WS. In: 2016 7th international conference on computer science and information technology (CSTT), 2016.

Ramteke J, Godhia D, Shah S, Shaikh A (2016) Election result prediction using Twitter sentiment analysis. In: International conference on inventive computation technologies (ICICT), 2016.

Ren Y, Wang R, Ji D (2016) A topic-enhanced word embedding for Twitter sentiment classification. Inf Sci 369:188–198

Sahu S, Rout SK, Mohanty D (2015) Twitter sentiment analysis a more enhanced way of classification and scoring. In: 2015 IEEE international symposium on nanoelectronic and information systems, pp 67–72

Saifa H, He Y, Fernandez M, Alani H (2016) Contextual semantics for sentiment analysis of Twitter. Inf Process Manage 52:5–19

Schumaker RP, Jarmoszko AT, Labedz CS Jr (2016) Predicting wins and spread in the Premier League using a sentiment analysis of twitter. Decis Support Syst 88:76–84

Shyamasundar LB, Jhansi RP (2016) Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms. In: India conference (INDICON), 2016 IEEE Annual

Siddiqua UA, Ahsan T, Chy AN (2016) Combining a rule-based classifier with weakly supervised learning for Twitter sentiment analysis. In: 2016 international conference on innovations in science, engineering and technology (ICISET)

Singh T, Kumari M (2016) Role of text pre-processing in Twitter sentiment analysis. Procedia Comput Sci 89:549–554

Smailovic J, Grcar M, Lavrac N, Znidarsic M (2014) Stream-based active learning for sentiment analysis in the financial domain. Inf Sci 285:181–203

Troussas C, Krouska A, Virvou M (2016) Evaluation of ensemble-based sentiment classifiers for Twitter Data. In: 2016 7th international conference on information, intelligence, systems & applications (IISA), 2016

Wang D, Al-Rubaie A (2015) Incremental learning with partial-supervision based on hierarchical dirichlet process and the application for document classification. Appl Soft Comput 33:250–262

Wang D, Al-Rubaie A, Clarke SS, Davies J (2017) Real-time traffic event detection from social media. ACM Trans Internet Technol 18(1):1–23

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush AM (2020) Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 EMNLP (systems demonstrations), pp 38–45, Nov. 16–20, 2020.

Wu F, Song Y, Huang Y (2016) Microblog sentiment classification with heterogeneous sentiment knowledge. Inf Sci 373:146–164

Yu Y, Wang X (2015) World Cup 2014 in the Twitter World: a big data analysis of sentiments in U.S. sports fans tweets. Comput Hum Behav 49:392–400

Yue L, Chen W, Li X, Zuo W, Yin M (2018) A survey of sentiment analysis in social media. Knowl Inf Syst 60:617–663

Zhao J, Gui X (2017) Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Transl Content Min 5:2870–2879

Zhao J (2016) Combing semantic and prior polarity features for boosting Twitter sentiment analysis using ensemble learning. In: 2016 IEEE first international conference on data science in cyberspace, pp 709–714, 2016.

Zhao J, Cao X (2015) Combining semantic and prior polarity for boosting twitter sentiment analysis. In: The 2015 IEEE international conference on smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015, pp 832–837, 2015

Zimbra D, Ghiassi M, Lee S (2016) Brand-related twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks. In: 2016 49th Hawaii international conference on system sciences, pp 1930–1938, 2016.