SentiFilter: A Personalized Filtering Model for Arabic Semi-Spam Content based on Sentimental and Behavioral Analysis

Mashael M. Alsulami[1], Arwa Yousef AL-Aama[2]
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Unwanted content in online social network services is a substantial issue that is continuously growing and negatively affecting the user-browsing experience. Current practices do not provide personalized solutions that meet each individual's needs and preferences. Therefore, there is a potential demand to provide each user with a personalized level of protection against what he/she perceives as unwanted content. Thus, this paper proposes a personalized filtering model, which we named SentiFilter. It is a hybrid model that combines both sentimental and behavioral factors to detect unwanted content for each user towards pre-defined topics. An experiment involving 80,098 Twitter messages from 32 users was conducted to evaluate the effectiveness of the SentiFilter model. The effectiveness was measured in terms of the consistency between the implicit feedback derived from the SentiFilter model towards five selected topics and the explicit feedback collected explicitly from participants towards the same topics. Results reveal that commenting behavior is more effective than liking behavior to detect unwanted content because of its high consistency with users' explicit feedback. Findings also indicate that sentiment of users' comments does not reflect users' perception of unwanted content. The results of implicit feedback derived from the SentiFilter model accurately agree with users' explicit feedback by the indication of the low statistical significance difference between the two sets. The proposed model is expected to provide an effective automated solution for filtering semi-spam content in favor of personalized preferences.**

***Keywords*—Personalization; sentiment analysis; behavioral analysis; spam detection; recommendation systems**

I. Introduction

Online Social Network (OSN) services provide online and instant communication in a large-scale manner. Despite the great social experience and communication benefits of these services, the vast usage of OSN services increases the amount of user-generated content, which brings several challenges and concerns regarding privacy, data management, information filtering, and content moderation. Users of such services are exposed to various kinds of content that can be unwanted or harmful [1].

Unwanted content can be defined as any electronic content, including text and multimedia that is not expected or welcomed by its final destination because of its disturbing or annoying nature. Unwanted content has been mostly considered and identified as spam content, which is received from undesired sources called spammers [2].

Solving the spam issue requires taking into consideration several aspects in order to propose solutions. These aspects include type of spam, where to detect spam, the form of spam, and how to detect it.

However, defining what spam is from users' personal perspectives needs further investigation. Personalization techniques could help to customize users' social space in OSN services and give the ability to recognize and detect what users really consider as spam messages to prevent them or block them from being received.

OSN services provide several interaction attributes that can be considered as indicators of users' perceptions of semi-spam content such as sharing/forwarding behavior, liking behavior, reporting behavior, and commenting behavior. Therefore, the authors of this paper assume that users use commenting behavior when they find a post they like or agree about or a post they do not want or disagree with. The aim of this work is to infer users' perception about a particular topic from detecting the sentiment of their comments to a post involving that topic combined with other behavioral factors.

In the context of our work, semi-spam content is defined as any electronic message that a particular user perceives as unwanted, unpleasant, annoying, or disturbing, based on his/her interaction behavior.

The work in this paper empirically assesses the impact of combining the sentimental factor of users' comments with liking behavior to detect semi-spam content for a particular user.

Accordingly, a personalized filtering model was designed and developed, which we named SentiFilter model, to filter out semi-spam content based on the sentiment polarity of users' comments combined with users' liking behavior. The effectiveness of using behavioral and sentimental factors in detecting semi-spam content was evaluated by comparing the implicit feedback derived from each behavioral and sentimental factor against users' explicit feedback about certain topics. More precisely, the work in this paper focuses mainly on Arabic messages, since there is very little research found in the literature on filtering Arabic spam content.

However, most of the existing studies concentrate on a particular definition of spam, such as inappropriate content, bullying content, racism, and hateful-speech content, without consideration of personalization or personal preferences. Another limitation in previous work is the focus on either sentiment or behavior as an indicator of users' preferences without making full use of both factors to detect semi-spam content for each individual. In this research, experimentation was carried out on the tweets dataset extracted from the timelines of 32 Twitter users to assess the impact of sentimental and behavioral factors in reflecting users' perceptions of a given topic. The research question that this paper aims to answer is as follows: Which behavioral or sentimental factors are more effective to detect semi-spam content in terms of the agreement between the implicit feedback, derived from each factor, and users' explicit feedback about a topic?

The main contributions of this paper are summarized as follows:

- Propose a personalized filtering model for semi-spam content, which we called SentiFilter that combines both sentimental and behavioral factors in detecting semi-spam content.
- Propose a personalized aggregate factorization algorithm, which we named the personalized aggregate factorization (PAF) algorithm that combines sentiment of users' comments with liking behavior to detect Arabic semi-spam content.
- Propose a list-based classifier that applies our proposed PAF algorithm to maintain users' blacklists, whitelists, and greylists.
- Compare the effectiveness of behavioral and sentimental factors in terms of the agreement between users' explicit feedback and the implicit feedback derived from the SentiFilter model.

The paper is organized as follows. Section II discusses the related work on personalized spam detection in the relevant literature. Then, an overview of the proposed SentiFilter model is demonstrated, including the proposed PAF algorithm, in Section III. Section IV explains the design of the experiment. Results and discussion are discussed in Section V. Finally, conclusion and future work are given in Section VI.

II. Related Work

In this section, we highlight previous work that proposed solutions to the problem of semi-spam messages from two points of view: spam detection solutions and personalization.

A. Spam Detection

Spam, as a term, has been used to define various types of unwanted content including spam emails, SMS spam, spam URLs, social spam, and web spam. Based on the definition of spam content, spam can be recognized by several forms such as malicious content, malware content, inappropriate content, not-safe-for-work content, or denial of service attacks [2]. Studies that involved users' perspectives in identifying spam content have used terms such as semi-spam [3] and grey spam [2]. Previous work in recommendation systems such as [4] and [5] benefited from the sentiment of customers' reviews to infer users' preferences and provide them with personalized recommendations.

Several interventions and techniques have been proposed in the literature to solve the problem of spam messages. Traditional techniques aim to block the source or distributor of inappropriate content while existing methods examine the content to extract features in order to recognize certain patterns and predict them using machine learning algorithms. Additionally, the definition of inappropriate content is varied in the literature. Some studies describe inappropriate content as spam content or not-safe-for-work content as in [6][7][8], where other studies focused on a specific type of content, such as text messages, and performed analysis processes to detect inappropriate content such as abusive language [9] or bullying behavior [10]. Spam URLs have also been investigated in [11] using behavioral analysis.

Based on existing literature in detecting unwanted content, this paper categorizes related work in this area of spam detection systems into list-based filtering and content-based filtering techniques. List-based spam detection methods aim to apply blacklisting techniques to detect the sources or the distributors of spam content and block them. On the other hand, content-based spam detection methods aim to extract features from the content itself to identify unwanted patterns.

1) List-based spam filtering: The concept behind list-based filtering is to create a blacklist of distributors of spam content. Those lists are blacklists, whitelists, and sometimes greylists. For instance, Tewari and Jangale [12] defined greylists by the sending pattern of the sender, where a greylist contains all unknown users who are initially rejected by the mail server. The mail server of the receiver will send a failure notification to the mail server of the sender. If the mail server of the sender sends the message again, the mail server of the receiver will accept it and move it from greylist to whitelist [12].

They classified senders by how many times they send the same message, relying on the fact that spam emails are usually sent in batches. On the other hand, Liu et al. [3] classified spam emails into two categories: complete spam and semi-spam emails. They considered complete spam emails as emails identified by all users as spam emails, while semi-spam emails are identified by crowdsourcing using trusted contacts. Therefore, a trust value needs to be assigned and computed for each contact [3].

O'Connor et al. [13] proposed a method to determine if a user is a source or a distributor of unwanted content by tracking his/her activities when a message was sent. They defined a set of metrics to decide if a certain user is a source of undesired electronic content. These metrics included message rate, block count, block rate, and message uniqueness. Their method can be considered as a collaborative method that collects information from users of a specific application to make a decision to ban or prevent users from that application or to add them to a watch list. The main goal of their method was to identify users who send such unwanted content. They defined patterns that indicated the distributor of

unwanted content because those users usually change their accounts' information periodically [13].

Bodkhe et al. [14] proposed a filtering method called Filter Wall (FW) to filter unwanted content based on a message trust management method. In their approach, a trust value was assigned for each message to classify it as wanted or unwanted. The trust value was assigned by users who were using the same application. The classification was based on computing the trustworthiness of each sender. The authors computed this value by aggregating the trust values for each message that each sender had sent and got its average [14].

Ma and Yan [15] also addressed the problem of blocking sources of unwanted content using a trust management method, which is a method in the communication field that is used to control unwanted traffic [15]. They proposed a system called PSNController to manage unwanted content. The brief concept of this system was to assign a trust value for each user. Users who used the same application had the ability to see the trust value of each other. PSNController is a customizable system to monitor unwanted content and identify its sources. They categorized unwanted content as bad text attacks, distributed denial of service, spammed multimedia, and viruses. They evaluated the system in terms of accuracy, efficiency, and robustness [15].

2) Content-based spam filtering: Content-based filtering (CBF) is mainly performed by determining the correlation between the content of items and user's preferences [16]. Applying CBF requires analyzing the content of each item to represent it as a set of features or terms, which is an expensive process.

Detecting spam messages based on their content is applied to several applications such as emotion recognition and inappropriate content detection tools. To detect spam messages based on their content, several approaches have been proposed in the literature.

In terms of Arabic spam detection, Mubarak and Darwish [9] studied the problem of detecting abusive Arabic text in social media. They used Twitter to create an Arabic corpus that contained a list of obscene words. They used that list to classify Twitter users based on their use of these words.

Another approach that applies CBF is proposed by Zitouni et al. [16]. They proposed a semantic content-based filtering technique, which benefits from the Web of Data concept, which is a term that refers to using all interconnected knowledge about different domains in the World Wide Web as a global database [15]. They integrated linked data with friend-of-a-friend (FOAF) vocabulary to enhance the semantics of data that are extracted from the web [16].

B. Personalization

Personalization can be defined according to [17] as the process of customizing content with respect to users' needs and preferences to enhance user experience. Personalization is the basic foundation of several studies including web content personalization [18], recommendation systems [19], and personalizing social media pages [20].

Personalization mainly depends on the construction of user profiles to get insights of what users may like or dislike. The information included in user profiles can be gathered from different sources such as log files and human resources indicating both implicit and explicit feedback [21].

Several studies have considered interaction attributes as implicit feedback of users' preferences. Bhavithra and Saradha [22] proposed a case-based reasoning strategy to recommend web pages based on the searching history of a particular user. They considered several interaction factors to be added in a user profile such as time on page, time on site, exit rate, and others. Their main aim was to benefit from these attributes to recognize patterns and apply collaborative filtering [22]. Moreover, Stai et al. [21] developed a mechanism to effectively personalize the enriched multimedia content based on users' interests and needs. They considered some interaction attributes to infer users' preferences such as "share video on social media, click on enrichment, click on ads, and playtime of a main video." Singh and Sharma [23] developed a multi-agent context-aware framework to personalize the web. They designed a dynamic user profiling technique to keep track of changes in users' behavior, which influences their interests.

Nabil et al. [24] proposed a sentiment-aware approach for article recommendation systems. They used consumers' reviews to detect feelings and infer preferences. They integrated both content-based and collaborative-based approaches to develop a hybrid recommendation system. Furthermore, a sentiment factor was considered by [25] to detect spammers. They found in their exploratory study that there were substantial differences between sentiments of spammers and sentiments of normal users. Hu et al. [25] incorporated a sentiment factor to a spammers detection framework to enhance the detection rate using sentiment analysis. A recent work by [26] proposed a protocol-based architecture model to predict direct and indirect interests of a user using the semantic relatedness concept.

III. An Overview of the Proposed SentiFilter Model

The proposed SentiFilter model was designed to utilize textual-based human emotional feedback through sentiment analysis to detect Arabic semi-spam content for a particular user. The results of the SentiFilter model are assessed through comparing the impact of liking behavior, commenting behavior, and the combination of liking behavior with sentiment of users' comments in effectively detecting semi-spam content for individuals.

To the best of our knowledge, the proposed SentiFilter model is the first to combine a sentimental factor of users' comments with other behavioral factors to detect semi-spam content for individuals. In our work, three factors are considered to model the proposed filter. Those are defined as follow:

- Liking behavior is a user's act that reflects liking reaction to a message, and it occurs when a user clicks on the like button.

- Commenting behavior is a reply act to a specific message.
- Sentiment factor represents the textual opinion/point of view of a user about a topic. It can be negative, positive, or neutral.

The SentiFilter model consists of three components and two classifiers. The three components are Extractor, Data preprocessor, and Detector. Each of the three components contains several functional modules. They are implemented as several python and R scripts. The two classifiers are a sentiment-based classifier, which is the core component in the SentiFilter model, and a list-based classifier, which is a data mining rule-based classifier. An overview of the structure of the SentiFilter model is shown in Fig. 1.

- Extractor: The extractor component is responsible for fetching and collecting implicit information about users' reactions to construct a user profile for each user. The Extractor consists of four modules. The workflow of the SentiFilter model starts by collecting the timeline of the incoming data stream from the user's social space, using the user's timeline Extractor module. Then, the user's comment Extractor module extracts all replies or comments from the user's timeline and prepares them to be passed to the data preprocessor component. All messages that a user has liked are collected, using the user's Like Extractor module. Each extracted comment is associated with its original message using the original message's Extractor module.
- Data Preprocessor: This component is responsible for cleaning Arabic text that was extracted, such as comments and tweets from the Extractor component. It involves removing meaningless and stop words.
- Arabic Topic Detector: This module is responsible for discovering the domain, subject, or topic that represents a particular message. Each domain *d* is represented by a set of Arabic keywords *T*, where $T_d = \{w_1, w_2, w_3, \ldots, w_n\}$. Arabic keywords for each topic are specified by crawling OSN messages of certain hashtags representing that domain. Then a preprocessing and tokenization over the collected set is carried out to extract the most frequent terms appearing in these messages. We propose to use hashtags, key phrases, and keywords to determine the topic of a particular message.
- Sentiment-based Classifier: This module is responsible for analyzing messages that are replies to other messages, with the goal to infer a user's attitude toward the original messages. The sentiment classification module is a predictive model that uses a supervised machine learning classification algorithm [27] to predict the polarity of a comment by examining its text.

The Term Frequency feature engineering method TF-IDF [28] was used to create a lexicon-based dictionary to identify positive, negative, and neutral keywords. This method aims at finding the most frequent words in a document for search purposes. The comments containing those keywords are manually labeled as positive, negative, or neutral messages. The labeled keywords are used by human annotators to train the sentiment classifier. The outcome of this module is a set of labeled messages that are passed to the next module. The workflow of this module is illustrated in Fig. 2.

- List-based Classifier: The Sentiment Classifier module and user's likes Extractor module work simultaneously to generate a rule-based classification model using a proposed personalized aggregate factorization (PAF) algorithm as shown in Fig. 3.

The proposed algorithm takes into consideration all previously mentioned factors to produce a personalized blacklist, whitelist, and greylist for each individual. There are seven cases in this algorithm, considering that I_m is the interaction behavior of a message m (i.e., a user likes a message), R_m is a reply to a message m, and d_m is a topic of a message m:

- Case 1: if I_m is null AND R_m is positive, THEN Whitelist (d_m).
- Case 2: if I_m is null AND R_m is neutral, THEN Greylist (d_m).
- Case 3: if I_m is null AND R_m is negative, THEN Blacklist (d_m).
- Case 4: if I_m is not null AND R_m is positive, THEN Whitelist (d_m).
- Case 5: if I_m is not null AND R_m is neutral, THEN Whitelist (d_m).
- Case 6: if I_m is not null AND R_m is negative, THEN Greylist (d_m).
- Case 7: if I_m is null AND R_m is null, THEN Blacklist (d_m).

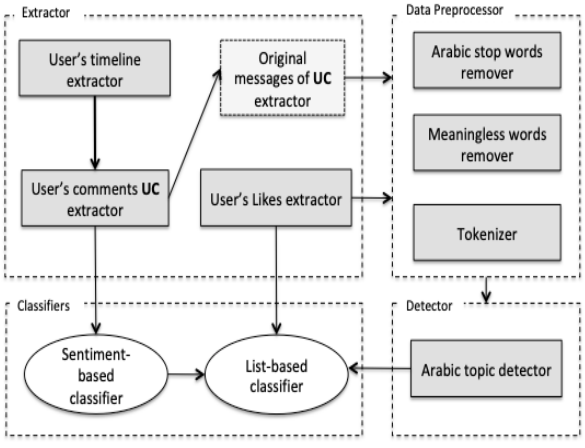

Fig. 1. The Structure of the Senti Filter Model.

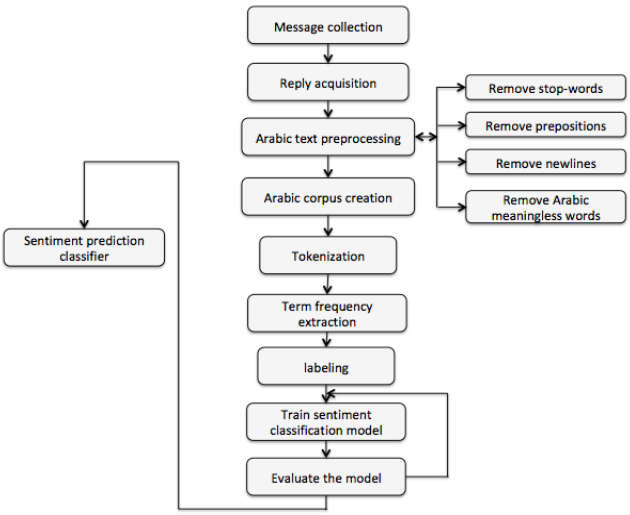

Fig. 2. The Workflow of the Sentiment-based Classifier.

Algorithm: Personalized Aggregate Factorization PAF algorithm

Input:
$\boldsymbol{I_m}$ is an interaction behavior of a message *m*, $\boldsymbol{R_m}$ is a reply to a message *m*, ***count*** is the number of times that a topic has been classified into a list, ***thres*** is a predefined threshold score that determines if a topic is definitely belongs to a particular list, ***sent*** is a sentiment polarity of a reply R for a message m, ***B*** is a blacklist for user *u*, ***W*** is a whitelist for user *u*, and ***G*** is a greylist for user *u*.
Output: classifying a message m into a list (labeling a message)
Procedure:

1. **for each** *m* in user social space **do**
2. **{**
3. Perform a topic detection for $m \rightarrow d_m$
4. **for all** d in user profile **do**
5. **if (**d_m is found**) then**
6. **if (**count (d_m) < thres**) then**
7. monitor()
8. **else** monitor()
9. **} // End for**

```
void monitor()
 {
  count(dm) ++
  if (Im is NULL AND Rm is Null) then
add dm to B
  else if (Im is not NULL AND Rm is Null) then
add dm to G
 else
  if (Im is not NULL AND Rm is not Null) then
  if (sent(Rm) is negative) then
add dm to G
  else  add dm to W
 else if (Im is NULL and Rm is not Null) then
if (sent(Rm) is positive) then
  add dm to W
else if (sent(Rm) is negative) then
  add dm to B
else add dm to G
 } // End of monitor
```

Fig. 3. The Proposed Personalized Aggregate Factorization Algorithm (PAF).

IV. Experiment Design

An experiment was conducted to empirically evaluate the impact of combining a sentimental factor with users' liking behavior to improve the filtering process of semi-spam content in terms of the agreement between the implicit feedback derived from the SentiFilter model and users' explicit feedback.

The experiment consisted of two main phases: 1) quantitative analysis of users' behavioral and sentimental factors, and 2) an analysis of users' explicit feedback towards certain topics using an online user survey instrument.

The aim of the first phase was to collect and analyze users' liking and commenting behaviors and to detect sentiment of users' comments. The objective of this phase was to examine the effectiveness of the proposed sentiment-based classifier by considering its accuracy using standard machine learning classification algorithms to select the most accurate one to be used in our analysis. In the second phase, a user survey was created asking users to rate several testing tweets to collect users' explicit feedback. The users' explicit feedback was used as a measurement to determine which behavioral and sentimental factors are closer to users' expectations.

A. Data Collection and Twitter API

Initially, a total of 80,098 Arabic tweets were collected using a Twitter API [29] for 32 users. The users in our sample were selected based on their number of tweets, comments, and likes to ensure their active status on Twitter. The mother language of all users was Arabic. They were selected from different backgrounds and interests. The gender distribution of the selected sample consisted of 21 females (65.6%) and 11 males (34.3%). The average number of comments for men was 52%, while it was 47.9% for women. The average number of likes for women was 51.75%, while it was 48.24% for men. The timeline of each user was crawled, including posts that he/she created or posts that were comments to other posts. Since the focus of this work is on personalization, we chose to collect timelines of users instead of collecting tweets for certain chosen topics and to select users who interacted with these topics as done in [4] and [20].

Furthermore, tweets that a user has liked were crawled in the same period of time. In order to demonstrate our methodology, Twitter is selected as a source of our datasets because Twitter API [29] is a freely available API for developers who wish to explore with real time data, unlike other OSN services such as WhatsApp messenger that have strict privacy constraints that will not enable us to collect users' conversations and their activities.

B. Sentiment Analysis

Inferring users' emotional feedback from their text-based comments is a critical task to determine individually semi-spam content. After collecting the 80,098 Arabic tweets, a reply acquisition process was performed to extract comments and exclude self-posts. We ended up with 6,307 comments with an average of 197.09 comments per user. Then, the following steps listed in the next subsections were performed:

1) Data preprocessing: Preprocessing Arabic text is a challenging task because of the sophisticated structure of the Arabic language. The SentiFilter model performed data preprocessing that included removing stop-words and Arabic prepositions. Meaningless Arabic words that did not add any meaning to the context of the text were removed. Some examples of those words are shown in Table I.

2) Tokenization: Basically, an Arabic corpus was created for each user representing his/her comments after the preprocessing phase. Then, each comment was tokenized to several tokens. A TF-IDF method was selected to extract terms that were used to label tweets. Human annotators were used to label tweets as positive, negative, or neutral.

3) Annotation and predictive classifier: The predictive classifier in this work is a sentiment-based classifier that uses supervised machine learning classification algorithms to predict the sentiment polarity of a comment and then produces a pair consisting of the sentiment polarity of the comment and the topic of its original tweet.

A support vector machine (SVM) supervised machine learning algorithm [30] was selected based on the performance analysis. It used to perform a multi-class classification and classify replies as positive, negative, or neutral.

In the training phase, we listed some terms that appeared in users' comments after the tokenization process and grouped them to three categories: positive, negative, and neutral. The two annotators were asked to use these categories to label comments.

Each reply was classified as positive, negative, or neutral based on the sentiment polarity of its words that were derived from the three categories.

C. 4.3. Arabic Topic Detection Model

This model is a rule-based model responsible for detecting the main topic for each extracted tweet. We defined five general topics, and for each topic, we created an Arabic domain-specific dictionary that contains the 20 most used keywords in that topic. We collected these keywords by crawling Arabic tweets using the Twitter API for hashtags related to that topic and filtered them manually. We then performed the same preprocessing and tokenization to get the 20 most frequent terms in each defined topic. Table II shows examples of keywords for each topic.

TABLE. I. EXAMPLES OF ARABIC MEANINGLESS WORDS

Arabic meaningless words	Transliterated words	Corresponding English words
من	mn	From
يا	Ya	Oh
على	3la	On
في	Fy	In/inside
ما	Ma	What
مع	M3	With
الف	alf	A thousand

TABLE. II. EXAMPLES OF KEYWORDS USED IN THE PROPOSED ARABIC TOPIC DETECTION MODEL

Topics (in English)	Topics (in Arabic)	Examples of keywords (in Arabic)	Examples of keywords (in English)
T1: Health	صحة	صحية – التهاب –طبية الاطباء – غذائية - الجراحي	i.e., (Health; medical; inflammation; surgery; your doctor)
T2: Sport	رياضة	الدوري – مباراة – الاتحاد – الاهلي – الهلال – دوري المحترفين	i.e., (League; match; (names of football teams); professional league)
T3 : Technology	تقنية	انترنت الاشياء – الذكاء الاصطناعي – تقنية – علم البيانات- تكنولوجيا	i.e., (IoT; AI; technology; data science; computer)
T4: Politics	سياسة	وزارة الداخلية – وزارة الخارجية – داعش – ايران - المرابطين	i.e., (Politics; Urdu; interior Minister; Minister of Foreign Affair; Nuclear power; armed forces)
T5: Social content	محتوى اجتماعي	مبروك – مبارك – شكرا – الشكر – عظم الله -	i.e., (Thanks, thank you, congrats; congratulation; you deserve it)

D. Interaction-based Detection

Interaction factors were defined by the mean of liking and commenting behaviors. After collecting users' timelines, 66,576 tweets that users liked with an average of 2,080.5 tweets per user were collected. Then, the same preprocessing steps to the text of these tweets were performed.

E. User Survey

An online user survey instrument was developed and used as a web-based application. The goal of the survey was to collect users' explicit feedback towards topics considered in the evaluation of the proposed filtering model.

The survey was distributed among the same 32 users who we considered in the first phase of the study. The survey consisted of two parts. In the first part, several tweets representing the five topics described in Table II were displayed and shown to each user as a Likert scale-based question. Each user was asked to rate each of the shown tweets based on how likely he/she would be to like/reply to/re-tweet them if they appeared in their social space. A scale from 1 to 5 was presented where 1 represented the not likely attitude, and 5 represented the extremely likely attitude. In the second part, users were asked to order the five topics based on their interests from the most interesting topics to them to their least-preferred topics. A user–topic factorization matrix of users' ratings was constructed to evaluate the effectiveness of the SentiFilter model in terms of the agreement of the implicit feedback derived from it with users' explicit feedback.

V. RESULTS AND DISCUSSION

The results obtained from the two phases of the experiment are demonstrated in two forms: the effectiveness of the sentiment-based classifier, and the comparison between the implicit feedbacks derived from of the SentiFilter model against the users' explicit feedback. In order to demonstrate our results, Table III shows statistical details about our datasets.

TABLE. III. DATASETS DETAILS

Number of crawled tweets	80,098
Number of users	32
Average number of tweets per user	2,503.063
Total number of comments	6,307
Average number of comments per user	197.0938
Number of likes	66,576
Average number of likes per user	2,080.5
Number of positive comments	610
Average number of positive per user	19.06
Number of negative comments	234
Average number of negative per user	7.3
Number of neutral comments	5,457
Average number of neutral per user	170.53
Number of intersected tweets between users' comments and likes	595
Average number of intersected tweets between users' comments and likes per user	18.59

The datasets distribution for the sentiment-based classifier is illustrated in Fig. 4, where each dataset represents a user.

The results from Fig. 4 revealed that most of the users' comments were neutral. Our definition of neutral comments is those comments that do not have any positive or negative keywords. The majority of neutral comments were basically replies such as personal congratulations, thanks, and good wishes. In the SentiFilter model, only 5% of the total comments were detected as negative comments, which indicates that negative comments are rarely posted in user social space in our datasets sample unless the content of the original tweet was negative.

We evaluated the effectiveness of the sentiment-based classifier using well-known performance measures for classification, which are accuracy, precision, and recall. Accuracy is a ratio of correct classified comments to the total number of comments of a user, while precision is a measure that determines how precise our model is in terms of the percentage of correct classified comments. On the other hand, recall is a ratio of total classified comments to total comments of a user [31]. An SVM classifier was selected as a base algorithm for our sentiment classifier because it produced the best results among other algorithms on average with an average accuracy of 90.89% as shown in Fig. 5.

From the statistical analysis of our collected datasets shown in Table III, the results indicated that 10.4% of liked tweets had positive comments and 87.5% of them had neutral comments, while only 2% of the liked tweets had negative comments. Thus, sentiment polarity of users' comments as a stand-alone factor cannot be used to reflect users' perception of semi-spam content, which means that negative comments do not indicate disliking attitude towards the topic under discussion. However, there is a relationship between commenting on a topic and detecting semi-spam content by the consideration of the silent reaction that was represented as case 7 in our proposed PAF algorithm.

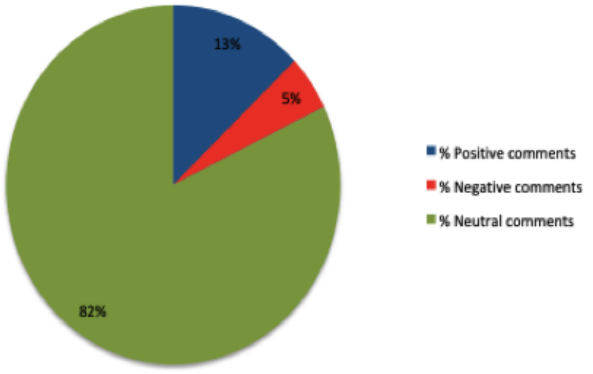

Fig. 4. The Datasets Distribution for the Sentiment-based Classifier.

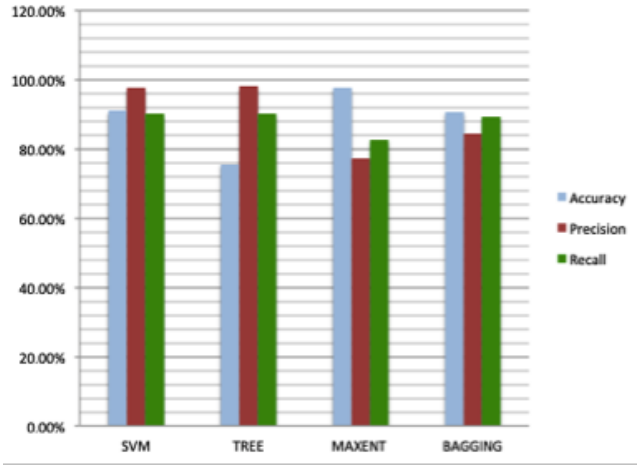

Fig. 5. The Classification Results of Sentiment-based Spam Classifiers.

The second phase of the evaluation was aimed to assess the effectiveness of the SentiFilter model in terms of the agreement between users' explicit feedback and the implicit feedback derived from behavioral and sentiment analysis. In this respect, a total of 30 users completed the survey. Users' explicit feedback was compared to the implicit feedback derived from liking behavior, commenting behavior, and the combination of sentimental factor with liking behavior by the mean indication. The statistical significance difference between each behavioral and sentiment factor with users' explicit feedback was computed and compared as shown in Table IV.

Results from Table IV show that there is no statistical significance difference between the implicit feedback derived from commenting behavior and users' explicit feedback (only 0.02 difference was found), which indicates that commenting behavior is the most effective behavioral indicator to significantly detect semi-spam content for an individual. On the other hand, the results of combining sentimental factor with liking behavior in the SentiFilter model are more effective to detect semi-spam content than considering liking behavior alone because the implicit feedback derived from the proposed model produces a high agreement with users' explicit feedback by the mean indication shown in Table IV.

TABLE. IV. THE STATISTICAL SIGNIFICANCE DIFFERENCES BETWEEN EACH FACTOR AND USERS' EXPLICIT FEEDBACK USING MEAN INDICATION (THE MEAN VALUE OF USERS' EXPLICIT FEEDBACK = 0.40)

Factor	Mean	Sig. Diff
Liking behavior	0.80	0.39
Commenting behavior	0.38	0.02
Liking behavior + Sentiment factor (SentiFilter model)	0.65	0.24

We measured the effectiveness of each factor by comparing the implicit feedback derived from each factor against users' explicit feedback by finding the significance difference between their mean values. The comparative analysis shown in Fig. 6 shows that there is only 0.24 statistical significance difference between implicit feedback derived from the SentiFilter model and users' explicit feedback, which indicates that the SentiFilter model accurately agrees with users' explicit feedback. However, implicit feedback derived from commenting behavior reports no statistical significance difference (only 0.02 difference) with users' explicit feedback.

The results also reveal that combining sentiment of users' comments with behavioral factors is a positive indicator to infer users' preferences, while negative comments do not reflect users' attitude towards semi-spam content.

The results and performance of the SentiFilter model might be varied if the knowledge is increased regarding the collected datasets.

Another constraint that can be encountered in the SentiFilter model is the social implication of using OSN services. We believe that users interact differently on different OSN services towards the same situations. For example, some users reply to others for the sake of courtesy in formal OSN services such as Twitter, while their reactions could be different if they were using an informal OSN service such as WhatsApp. Therefore, the SentiFilter model might report different findings when another OSN service is considered.

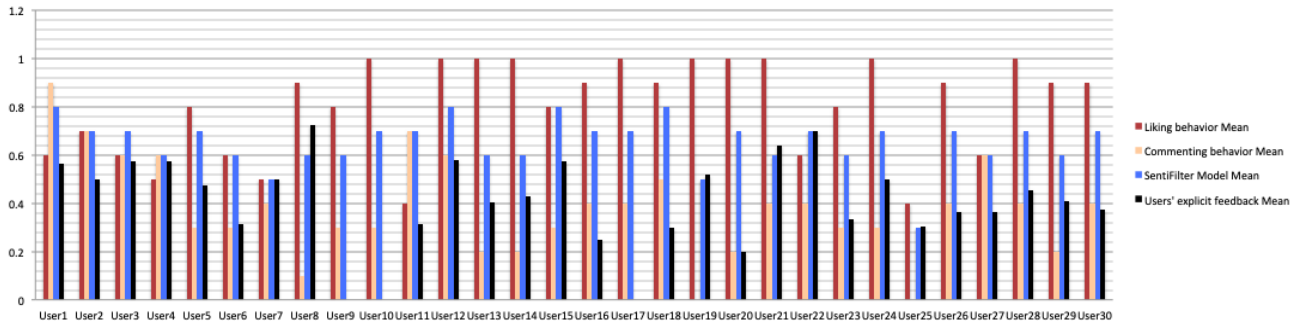

Fig. 6. Comparative Analysis of Sentiment and behavioral Factors with users' Explicit Feedback.

VI. CONCLUSION AND FUTURE WORKS

Personalization is a fundamental task in most aspects of our daily life. It provides users with a better experience and gives them more control over what they really want. Employing this concept in spam detection systems can help in enhancing users' browsing experience by automatically generating personalized filters that are able to recognize and control what users see or receive in their social spaces. Thus, this paper introduces a new factor that can be combined with other behavioral factors to be used as an indicator to detect semi-spam messages, which is the sentiment factor. The proposed SentiFilter model empirically assesses the impact of combining the sentimental factor with behavioral factors to filter semi-spam messages. Our results showed that commenting behavior is more effective than other behavioral factors in detecting semi-spam content by the high agreement between the implicit feedback derived from it and users' explicit feedback.

We have evaluated the effectiveness of the SentiFilter model in terms of the agreement between its implicit feedback and users' explicit feedback. The implicit feedback derived from the SentiFilter model accurately agrees with users' explicit feedback by the indication of the statistical significance difference between the two sets.

In our future work, we will concentrate on overcoming the knowledge limitation by increasing the sample size to produce more general results. We will apply the same model to English messages to find out if any different observations can be drawn. We will also work on designing and developing blocking mechanisms that maintain users' blacklists in effective ways. Furthermore, we will plan to identify some platform-specific behaviors that differentiate how users perceive semi-spam content. We encourage researchers in both HCI and information security fields to incorporate our findings in their design decisions to effectively maintain users' blacklists.

REFERENCES

[1] M. M. Alsulami and A. Y. Al-Aama, "Exploring User's Perception of Storage Management Features in Instant Messaging Applications: A Case on WhatsApp Messenger," Proc. 2019 2nd Int. Conf. Comput. Appl. Inf. Secur., pp. 1–6.

[2] K. S. Bajaj, "A multi-layer model to detect spam email at client side," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng., vol. 198 LNICST, pp. 334–349, 2017.

[3] X. Liu et al., "CPSFS: A credible personalized spam filtering scheme by crowdsourcing," Wirel. Commun. Mob. Comput., vol. 2017, 2017.

[4] R. Harakawa, D. Takehara, T. Ogawa, and M. Haseyama, "Sentiment-aware personalized tweet recommendation through multimodal FFM," Multimed. Tools Appl., pp. 18741–18759, 2018.

[5] M. M. Alsulami and R. Mehmood, "Sentiment Analysis Model for Arabic Tweets to Detect Users' Opinions about Government Services in

Saudi Arabia: Ministry of Education as a case study," ALYamamah Inf. Commun. Technol. Conf.

[6] B. K. Narayanan, R. B. M, S. M. J, and M. Nirmala, "Adult content filtering : Restricting minor audience from accessing inappropriate internet content," Educ Inf Technol, 2018.

[7] G. Karyono, A. Ahmad, and S. A. Asmai, "Survey on Nudity Detection : Opportunities and Challenges based on ' Awrah Concept in Islamic Shari ' A," J Ther Appl Inf Technolo, vol. 95, no. 15, pp. 3450–3460, 2017.

[8] F. Aiwan and Y. Zhaofeng, "Image spam filtering using convolutional neural networks," Pers Ubiquit Comput, no. 22, pp. 5–6, 2018.

[9] H. Mubarak and K. Darwish, "Abusive Language Detection on Arabic Social Media," Proc. First Work. Abus. Lang. Online, pp. 52–56, 2017.

[10] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," Proc. 2017 ACM Web Sci. Conf., pp. 13–22.

[11] C. Cao and J. Caverlee, "Detecting Spam URLs in Social Media via Behavioral Analysis," Eur. Conf. Inf. Retr., pp. 703–714.

[12] A. Tewari and S. Jangale, "Spam Filtering Methods and machine Learning Algorithm - A Survey," Int. J. Comput. Appl., vol. 154, no. 6, pp. 8–12, 2016.

[13] B. D. O'Connor, "System And Method For Detecting Unwanted Content," US9948588B2, 2018.

[14] T. G. and V. J. Renushree Bodkhe, "A Novel Methodology to Filter Out Unwanted Messages from OSN User's Wall Using Trust Value Calculation," Proc. Second Int. Conf. Comput. Commun. Technol., pp. 755–764, 2016.

[15] Z. Yan, "PSNController- An Unwanted Content Control System in Pervasive Social Networking Based on Trust Management," ACM Trans Multimed Comput Commun Appl, vol. 12, no. 1, p. 17, 2015.

[16] H. Zitouni, S. Meshoul, and K. Taouche, Enhancing Content Based Filtering Using Web of Data, vol. 1. Springer International Publishing, 2018.

[17] P. Germanakos and M. Belk, Human- Centred Web Adaptation and Personalization From Theory to Practice. Switzerland: Springer International Publishing, 2016.

[18] S. Ferretti, S. Mirri, C. Prandi, and P. Salomoni, "The Journal of Systems and Software Automatic web content personalization through reinforcement learning," J. Syst. Softw., vol. 121, pp. 157–169, 2016.

[19] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "iSCUR : Interest and Sentiment-Based Community Detection for User Recommendation on Twitter," Switz. Springer Int. Publ., pp. 314–319, 2014.

[20] U. K. Wiil, "Emotion-Based Content Personalization in Social Networks," vol. 42, no. 1, pp. 1–16, 2018.

[21] E. Stai, S. Kafetzoglou, E. E. Tsiropoulou, and S. Papavassiliou, "A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content," Multimed. Tools Appl., vol. 77, no. 1, pp. 283–326, 2018.

[22] J. Bhavithra and A. Saradha, "Personalized web page recommendation using case-based clustering and weighted association rule mining," Cluster Comput., vol. 0123456789, pp. 1–12, 2018.

[23] A. Singh and A. Sharma, A Multi-agent Framework for Context-Aware Dynamic User Pro fi ling for Web Personalization. Springer Nature Singapore Pte Ltd. 2019.

[24] S. Nabil, J. Elbouhdidi, and M. Yassin, "Recommendation system based on data analysis- Application on tweets sentiment analysis," 2018 IEEE 5th Int. Congr. Inf. Sci. Technol., pp. 155–160, 2018.

[25] X. Hu, J. Tang, H. Gao, and H. Liu, "Social Spammer Detection with Sentiment Information," IEEE Access, 2014.

[26] S. Goel and R. Kumar, "SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction," Expert Syst. Appl., vol. 141, p. 112955, 2020.

[27] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, 2019.

[28] J. Guo, Y. Mu, M. Xiong, Y. Liu, J. Gu, and J. Garcia-Rodriguez, "Activity Feature Solving Based on TF-IDF for Activity Recognition in Smart Homes," Complexity, vol. 2019, 2019.

[29] "TwitterAPI Documentation," 2017.

[30] P. W. Wang and C. J. Lin, "Support vector machines," Data Classif. Algorithms Appl., no. ii, pp. 187–204, 2014.

[31] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," Adv. Neural Inf. Process. Syst., vol. 2018-Decem, no. NeurIPS, pp. 1920–1930, 2018.