Classification of Facebook News Feeds and Sentiment Analysis

Shankar Setty*, Rajendra Jadi†, Sabya Shaikh‡, Chandan Mattikalli§, Uma Mudenagudi**
B. V. Bhoomaraddi College of Engineering and Technology, Hubli-India
{shankarsetty*, rajendrarjadi†, sabya.shaikh‡, mattikallichandan§, uma.mudenagudi**}@gmail.com

***Abstract*—As recently seen in Google's Gmail, the messages in inbox are classified into primary, social and promotions, which makes it easy for the users to differentiate the messages which they are looking for from the bulk of messages. Similarly, a users wall in facebook is usually flooded with huge amount of data which makes it annoying for the users to view the important news feeds among the rest. Thus we aim to focuses on classification of facebook news feeds.**

In this paper, we attempt to classify the users news feeds into various categories using classifiers to provide a better representation of data on users wall. News feeds collected from facebook are dynamically classified into various classes such as *friends posts* and *liked pages posts*. *Friends posts* are further categorized into *life events posts* and *entertainment posts*. Posts or updates from pages which are liked by the users are grouped as *liked pages posts*. Posts from friends are tagged as *friends posts* and those regarding the events occurring in their lives are said to be *life event posts* and the rest are tagged as *entertainment posts*. This helps users to find "important news feeds" from "live news feeds". Sentiments are important as they depict the opinions and expressions of the user. Hence, detecting the sentiments of users from the *life event posts* also becomes an essential task. We also propose a system for automatic detection of sentiments from the *life event posts* and categorize based on sentiments into happy, neutral and bad feelings posts. This paper looks towards applying the classification methods from the literature to our dataset with the objective of evaluating methods of automatic news feeds classification and sentiment analysis which in future can provide facebook page a well organized and more appealing look.

***Keywords*—Facebook news feeds, Text classification, Sentiment analysis.**

I. Introduction

With the drastic increase in the data on social networking websites it has now become important to put some structure to these data for users to have pleasant experience. Due to the easy pattern and format of these websites the accessing rate is growing exponentially. Facebook [1] users posts their views, opinions and their point of perception on different topic. It may include political issue, religious issue, technology, product, movie review and much more daily gossiping issues flooded in their surroundings. Usage of social networking sites like Facebook, Twitter, Myspace, Google+, LinkedIn has shown a rapid increase over years. It signifies that users have moved from traditional trends like mail, blog to micro-blogging and social networking sites. Today, users of social networking sites are more interested in friends life status, rather than posts about product updates from companies, etc for which they subscribe or follow different websites. Users are currently forced to visit multiple sources and scan through various contents before finding the useful information. For such a long-term information need, the best way to help users is to classify the data based on importance of information.

Fig. 1. Classified Facebook Data: Column 1 shows liked pages posts, column 2 shows friends posts which are subcategorized into life posts and entertainment posts. Sentiments such as happy, neutral and sad are displayed for life event posts.

In this paper, we propose a system for classification of facebook news feeds and automatic detection of sentiments. Gmail [2] has been enriched with automatic classification of mails into primary, social and promotions. This has benefited users to find important mails. We adopted the approach of gmail to automatically classify facebook news feeds into life posts and entertainment posts. Further we perform sentiment analysis of life event posts. This provides a better structure to the facebook. Figure 1 shows output of our proposed system for classification of facebook news feeds and sentiment analysis.

In our work we considered facebook for following reasons. Firstly, facebook is well known and frequently accessed site across the globe. According to eMarketer [3], a digital marketing analysis firm, facebook is still the number one social

Fig. 2. Comparison of facebook categories (first column) and sentiment allotment (second column) in existing facebook (manual selection) and our approach (automatic allotment).

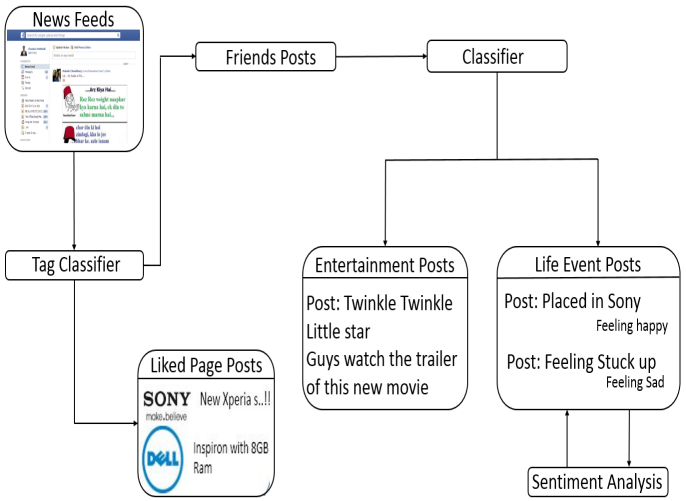

Fig. 3. An illustration of flow of our proposed classification system

network by a large margin. Secondly, facebook is not biased to any particular users, category or crowd. Facebook belongs to general public whose opinions are really worthwhile for any general survey. And lastly, facebook is used by many people from different countries.

Usage statistics obtained from facebook provides a rough estimate that the typical facebook user receives over 1,000 posts per week from nearly 130 friends [3]. Although a user spends around 55 minutes per day on the site [3], users are likely to miss some potentially interesting content. This highlights the need for better tools to surface the most important news feeds. Furthermore, not much has been implemented in facebook for classification to find the most important news feeds. Also there is no functionality for automatic detection of sentiments of a post. Users have to explicitly select the sentiments to their posts provided by facebook as shown in Figure 2. The left side of the Figure 2 shows comparison of manual labeling life event posts in facebook and our approach which categorizes posts automatically. The right side of the Figure 2 shows manual selection of sentiments in facebook and automatic detection of sentiments in our approach.

The contribution of the proposed system is as follows:

1) An approach to structure the facebook news feeds is provided.
2) A tool is developed to automatically pull live news feeds of facebook user. No human intervention needed to build the corpus.
3) A classifier is developed which classifies fresh data based on training data.
4) Sentiment analysis system has been developed to detect the sentiments of the live news feeds.
5) Experimental evaluations have been conducted to produce results on facebook live news feeds.

The live news feeds is collected from facebook using API's provided by facebook developer to extract data. The posts collected from facebook is dynamically splitted into liked pages posts and friends posts as showed in Figure 3. The friends posts are then categorized as life event posts and entertainment posts and further life event posts are used to evaluate the sentiments of users through the corpus collected from posts of users' wall as shown in Figure 3. Our approach provides a valuable input to facebook to extend its features for classifying its post into various categories.

The rest of the paper is organized as follows. Section II presents the related work. Section III deals with methodology used for classification of facebook news feeds. Experimental results are discussed in Section IV. Conclusion and future work is dealt in Section V.

II. Related Work

In this section, we give a brief survey of proposed approaches for text classification and sentiment analysis in literature. In [4] authors speak about opinion mining on facebook comments. Around 2000 comments from facebook were collected which were splitted automatically into three sets as comments containing positive impact such as good, best, happy, comments containing negative impact bad, worst, sorrow and comments containing average impact neutral, average, fine. Largely in paper [4] only the opinion of the user regarding a specific topic is analyzed. The focus was on opinion extraction and classification of realtime facebook status.

In [5] authors demonstrate that although the majority of company posts on facebook are aimed for direct sales and promotions, it is their communication messages that received the most attention from customers. Being able to identify the type of messages sent on facebook, and measure the attention of those messages received from the facebook fans, may assist managers in developing the right social-media marketing strategies. Direct marketing and communication messages have different communication goals, and thus can be considered as two different genres and these two message types differ in the topics what paper [5] considered. On these lines, in our approach we considered facebook live news feeds to be classified into life posts and entertainment posts.

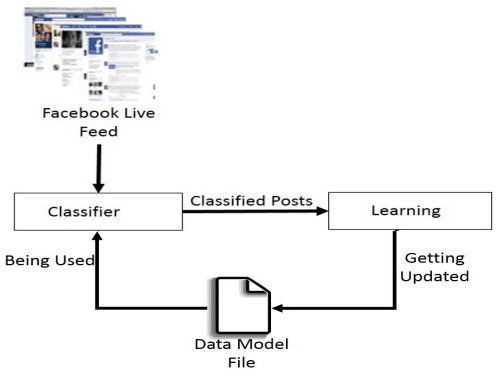

Fig. 4. Model showing automated training dataset

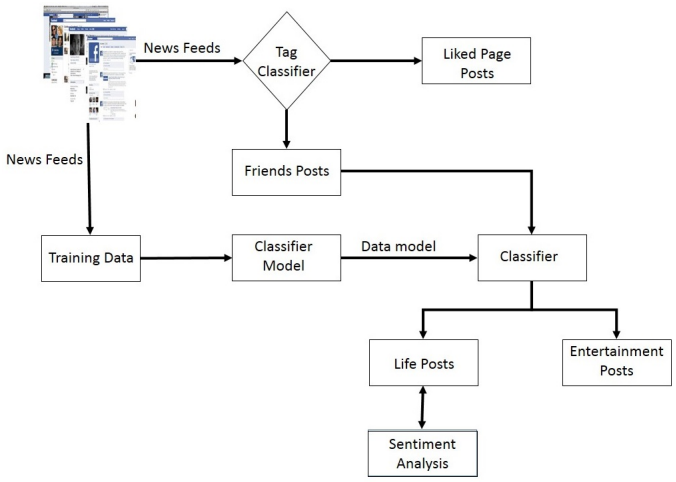

Fig. 5. Model showing classification and sentiment analysis of facebook news feeds

In [6] authors task is characterized by attempting to determine age/child appropriateness, despite the presence of unbalanced class sizes and the lack of quality training data with human judgements of appropriateness. The paper explores the classification problem for news feeds determining whether short snippets for individual feed entries are appropriate for children. The problem needed features beyond the term space, like the text readability features. In our work we consider classification of facebook news feeds.

In [7] authors focus is on classifying sentiment of facebook status updates using binary and multi-class labels. Facebook makes a distinction between a facebook users status update, versus links to a news article or other source of information, versus comments that are a response to another facebook user. Unlike tweets, status updates can use upto 420 characters. Thus, status updates more often are written in mostly sentence-like structures that can benefit from Parts Of Speech (POS) analysis. For binary classification, positive and negative sentiment labels were chosen for classification. These labels were represented as POSITIVE: smiley, wink, tongue, angel, shades, blush, rock-on and NEGATIVE: frown, shock, skeptical, evil, angry, fail. For the multi-class case, four sentiment labels: unhappy, happy, skeptical and playful were chosen to maximize the amount of usable data. These labels are represented by UNHAPPY: evil, frown, shock, angry, HAPPY: smiley, SKEPTICAL: skeptical, PLAYFUL: angel, rock-on, shades, tongue, wink. We try to adopt a similar idea to automatically categorize sentiments for facebook news feeds.

However, based on our survey there is no previous work carried on classification of facebook live news feeds into life events posts and entertainment posts. And automatic detection of sentiments from life events posts and categorization into happy, neutral and bad.

III. Methodology

A. Automated training dataset

Around 2000 manual labeled posts are used to create data model for initial classification. As shown in Figure 4 the classifier classifies the posts into appropriate classes (liked page post, friends posts - life posts and entertainment posts) by learning to create a data model. The automated training dataset, automatically collects the live news feeds from facebook and classifies the posts automatically into appropriate classes to update the existing data model.

B. Our Approach

In this section, we devised an approach to classify facebook news feeds and perform sentiment analysis. As shown in Figure 5, firstly facebook news feeds are fetched from the facebook. The tag classifier is built which classifies facebook news feeds based on the category tag of a post into liked pages posts and friends posts. A learning based classifier is built using various classification algorithms such as Binary Logistic Regression [8], Naive Bayes [9], Support Vector Machine (SVM) [10], Bayes Net [11] and J48 [12], which further classifies friends posts into life events posts and entertainment posts. The trained dataset is created using automated training dataset as mentioned in Section III-A. The training dataset is fed to classifier for learning. The test set (new posts) is classified based on learnt classifier into life events posts and entertainment posts. Further, life events posts are tagged into happy, neutral and sad posts based on sentiment score value determined using POS tagger and SentiWordNet dictionary [13].

The steps followed in our approach are:

1) Data Collection and Feature Extraction: The data is extracted from the facebook using facebook Restfb Java APIs [14]. Around 2000 posts are collected from different users which is considered as training set. Live news feeds are used as test data to classify using the different classifiers.

2) Classification: The various classes considered are as follows:

- **Liked pages posts**: Most of the users like various companies pages. These companies periodically update their status which is subsequently pushed to user's wall.
- **Friends posts**: Posts or status updates that are mainly contributed by user's friends or followers contribute to the friends posts and these are subsequently pushed to

the news feeds of user. The friends posts are further classified into life events posts and entertainment posts classes.

- **Life events posts**: User posts about the events occurring in their lives like engagement, marriage or their present status like traveling, having fun, etc. All these are considered as life events posts.
- **Entertainment posts**: Updates like poems, reviews about movies, technologies and products, etc which are not of primary importance are labeled as entertainment posts.

In classification, training set is used to build a model that can classify the data samples into known classes. The classification process involves following steps:

- The classification of liked pages posts and friends posts is done using "category tag" of the posts.
- Classifiers are used for classification of life and entertainment posts. A training data set consisting of 2000 posts was created which was manually labeled and a model was created.
- The built model was used to classify the live news feeds.

The various types of classification techniques used to build classifier are:

Binary Logistic Regression [8] is a type of regression analysis where the dependent variable is either 0 or 1. The logistic regression model is simply a non-linear transformation of the linear regression. The logistic distribution is an S-shaped distribution function where the estimated probabilities lie between 0 and 1. The estimated probability suggests the probability that the posts belongs to a particular class. If the probability is below 0.5 than it is class 0 and if it is above or equal to 0.5 it is class 1.

Naive Bayes Classifier [9] is probabilistic model which implements Bayes theorem with strong independence assumptions. The probability of each post belonging to different classes are computed. The post belongs to the class with highest probability.

SVM [10] Support Vector Machine is supervised learning model. It constructs a set of hyperplanes in a high or infinite dimensional space which is used for classification. A good classification is achieved if the hyperplane has the largest distance to the nearest training data point.

$$h = wx + b; \tag{1}$$

In equation 1, w is the scalar weight of the post, b is bias to move a hyperplane and h in a particular direction. Equation 1 is used to classify a post into a class 1 or class 2 based on the h (hyperplane value).

Bayes Net [11] is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.

J48 [12] The decision trees generated by J48 is used for classification. J48 builds decision trees from a set of training data samples. A decision tree is a predictive model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

3) Sentiment Analysis: Pre-processing of news feeds is done by removing of special characters and stemming of words. For example, the stem of "connections", "connecting" and "connected" is "connect". The complexity of the text is reduced by pre-processing. The text is then summarized as dichotomous variables, indicating the presence or absence of each stem. If the stem is present in the SentiWordNet dictionary the polarity of the word is returned. Sentiment analysis has different meanings in different contexts, in our context it is defined as "the task of identifying polarity which suggests emotions of the posts".

In our approach, the system developed for sentiment analysis splits the input text into words and then,

- Tokenization is performed on facebook posts to extract each term.
- Using POS tagger, part-of-speech of each term is detected.
- The input facebook post is split into words. The polarity of each word is found using SentiWordNet dictionary.
- To filter the terms for scoring, stopwords are removed and SentiWordNet dictionary is used to determine sentiment orientation value of a word.
- The sentiment score is calculated for each word. Each word has a sentiment orientation value, such as +1(strongly positive), +0.5(weakly positive), -1(strongly negative), -0.5(weakly negative). The sentiment score of each word is calculated and their sentiment scores are summed-up. The resultant sentiment score is used to detect sentiment of the post and thus user.
- Sentiment scores range from [-1,+1].
- For terms with multiple polarity scores, the average of all these scores is used.
- Terms not found in the SentiWordNet dictionary where regarded as non-sentiment bearing words.

Our approach of combining classification and sentiment analysis is unique of its kind for facebook application.

IV. EXPERIMENTS AND RESULTS

In this section, we first present the details of our experimental procedure including descriptions of the data used, classification using Weka APIs [15] and learning model, sentiment scoring. We then evaluate and analyze experimental results.

A. Data Used and Corpus Analysis

In our approach facebook news feeds are used which are primary focus for classification. These news feeds will also be used to detect the sentiments on the basis of features contained

TABLE I. ACCURACY OF CLASSIFICATION ALGORITHMS

Algorithms	Correctly classified instances	Kappa statistic
BayesNet	94.69%	0.89
J48	93.73%	0.92
NaiveBayes	92.70%	0.86
SVM	94.75%	0.86
Logistic Regression	92.71%	0.86

in the news feeds extracted. The Restfb APIs is used to extract data from facebook. News feeds of one to two weeks are extracted and stored from November 2013 to January 2014. The automated training data set is considered for labeling of posts from news feeds into life and entertainment categories. This involves manually labeling a number of posts with the chosen categories. Among these, $2/3^{rd}$ posts were considered for training data and the rest of the posts as test data.

B. Classification

The classification of posts into friends posts and liked pages posts is done using "category tag" of posts in facebook news feeds. The classification of friends posts into life events posts and entertainment posts is done using two techniques: (i) Weka APIs [15] and (ii) Learning Model for Classification.

1) Using Weka APIs: An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. All the entries in a feature vector array are converted into Attribute Relation File Format (ARFF) [16] to serve as input for the classifier. This input is used to train the classifier. Once it is trained, different classifiers like Binary logistic, Naive Bayes, SVM, Bayes net and J48 are used for classification of life events posts and entertainment posts.

2) Learning Model for Classification: We implemented SVM and logistic regression learning models for comparing with Weka APIs based classifiers. SVM and logistic regression learning model is created in the form of hypothesis and this hypothesis is applied to classify the posts. Features such as punctuation marks, new lines, words, phrases, emoticons, names of people appearing in posts and the caption field of the posts are considered. From analysis of the data obtained in our experimentation we observed features such as punctuation marks, posts with greater than 3 lines of text, 'tongue out' emoticons in the post belong to entertainment posts and features such as friends name in the post, short text posts belong to life event posts. The total weights of posts that are calculated from these features are given as inputs to logistic regression model and SVM model.
The hypothesis for logistic regression used is:

$$h(x) = \frac{1}{1 + e^{\theta^T X + b}} \tag{2}$$

In equation 2, θ is the array of coefficients whose length is equal to number of features, X is the array of features for all the posts and b is the intercept. Gradient descent function is used to get a minimized set of values for θ that reduces the cost and provides a better classification hypothesis. $h(x)$ is the probability. If the probability is greater than or equal to 0.5 it belongs to life events category otherwise entertainment category.

TABLE II. COMPARISON OF VARIOUS PARAMETERS FOR EACH CLASSIFICATION ALGORITHM

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BayesNet	0.93	0.07	0.93	0.93	0.93	0.97
J48	0.96	0.04	0.96	0.96	0.96	0.97
NaiveBayes	0.93	0.07	0.93	0.93	0.93	0.97
SVM	0.95	0.05	0.95	0.95	0.95	0.98

TABLE III. COMPARISON OF ACCURACY USING WEKA AND LEARNING MODEL (SVM & LOGISTIC REGRESSION) - TRAINING SET I

No. of posts	Weka's SVM	SVM Learning Model	Weka's Logistic Regression	Logistic Regression Learning Model
200	93.60%	98.50%	93.01%	93.50%
400	93.69%	98.50%	92.70%	91.50%
600	93.46%	97.66%	91.10%	86.50%
800	93.51%	97.75%	91.93%	85.60%
1000	93.47%	97.13%	91.31%	83.85%

C. SentiWordNet Scoring

Each news feed is parsed by SentiWordNet to determine the polarity. The SentiWordNet dictionary requires a term and its part-of-speech to produce a sentiment score. We depict sadness with negative score, happiness with positive score and terms that are not found in the SentiWordNet dictionary are automatically scored as neutral. All terms in each post are represented by term and its sentiment score. In situations where a term had more than one score, the average score is chosen. Sentiment score of each term of a post is summed to find the polarity of the post. Based on the polarity of the post we categorize into "Happy", "Neutral" and "Sad".

D. Experimental Results

Accuracy of classification of posts is evaluated for different classifiers using Weka and also our learning model approach. Two fold cross validation is used to evaluate the accuracy using Weka. The various classification algorithms used for comparison are Bayes Net, J48, Naive Bayes, SVM. Approximately 2000 posts are considered for analysis. The accuracy obtained for each algorithm is listed in Table I. According to the results shown in Table I, SVM followed by Bayes Net shows better accuracy than other classification algorithms for our dataset.

True Positive (TP) rate, False Positive (FP) rate, Precision, Recall, F- Measure and ROC are some parameters on the basis of which we evaluated the performance of the classifiers as shown in Table II. SVM has shown consistent performance for all the parameters. Hence SVM is comparatively better than other algorithms for our dataset.

We collected a dataset of facebook news feeds of approximately 2000 posts from 3 different time period scenarios. (i) Training Set I: collected dataset in the second week of November 2013. (ii) Training Set II: collected dataset in the second and third week of December 2013. (iii) Training Set III: collected dataset in the second and third week of January 2014.

Accuracy of classification of posts using logistic regression and SVM implemented using Weka and learning model are evaluated using hold-out method [17]. For the sake of performance analysis of weka versus learning model we considered

TABLE IV. COMPARISON OF ACCURACY USING WEKA AND LEARNING MODEL (SVM & LOGISTIC REGRESSION) - TRAINING SET II

No. of posts	Weka's SVM	SVM Learning Model	Weka's Logistic Regression	Logistic Regression Learning Model
200	93.44%	98.5%	89.94%	71.5%
400	93.22%	97.75%	91.97%	73.5%
600	93.00%	97.16%	90.98%	68.83%
800	92.96%	96.5%	92.99%	70.35%
1000	92.96%	96.1%	92.49%	73.4%

TABLE V. COMPARISON OF ACCURACY USING WEKA AND LEARNING MODEL (SVM & LOGISTIC REGRESSION) - TRAINING SET III

No. of posts	Weka's SVM	SVM Learning Model	Weka's Logistic Regression	Logistic Regression Learning Model
200	92.84%	97.48%	92.49%	64.82%
400	92.97%	96.49%	95.09%	60.65%
600	92.85%	96.82%	91.78%	62.93%
800	93.37%	96.37%	91.48%	64.20%
1000	93.68%	96.6%	92.24%	63.8%

experimentation on 3 different time peroids facebook datasets as shown in Table III,Table IV and Table V.

As shown in Table III, Table IV, Table V, SVM based learning model's performance is better than weka SVM and weka logistic regression's performance is better than logistic regression learning model. The accuracy obtained for learning model for identified features with SVM is in the range of 97% to 99% as compared to weka SVM for varying size of posts from 200 to 1000. Similarly accuracy obtained for learning model for identified features with logistic regression is in the range of 60% to 93% as compared to weka logistic regression for varying size of posts from 200 to 1000. The accuracy seems to be very high as the features learnt in the learning model have influenced the performance as compared to weka. However, the accuracy could vary if the datasets are changed overtime.

It is observed that learning model has outperformed weka based on the features learnt on the facebook dataset used. However, the performance could vary based on the facebook dataset for both weka and learning model.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a system for classification of facebook news feeds. We developed a model to classify posts appearing on users facebook wall to find most important news feeds and to automatically detect the sentiments of the user. The approach to structure the data and detect the sentiments of users in facebook reduces manual survey work which is done for drawing conclusions on opinions posted on facebook. Experiments on the live news feeds showed that the proposed approach could achieve significantly improved performance for structuring the data on facebook using SVM classifier learning model.

There are several avenues for future work arising from our work. The proposed work could further be extended to other social networking websites and can be applied to other types of social media data (like customer reviews, blog messages, comments).

REFERENCES

[1] Facebook developers. [Online]. Available: https://developers.facebook.com/

[2] Google inbox categories. [Online]. Available: https://support.google.com/mail/answer/3094596?hl=en

[3] Emarketer. [Online]. Available: http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976

[4] A. Shrivatava and B. Pant, "Opinion extraction and classification of realtime facebook status," *Global Journal of Computer Science and Technology*, vol. 12, pp. 35–40, 2012.

[5] B. Yu and L. Kwok, "Classifying business marketing messages on facebook," in *Proceedings of the SIGIR 2011 Workshop on Internet Advertisement*, 2011.

[6] T. Polajnar, R. Glassey, and L. Azzopardi, "Detection of news feeds items appropriate for children." in *ECIR*, ser. Lecture Notes in Computer Science. Springer, 2012.

[7] J. K. Ahkter and S. Soria, "Sentiment analysis: Facebook status messages," The Stanford NLP Group, Stanford University, Natural Language Processing, Final Project Report, 2010.

[8] Y. Cheng, K. Zhang, Y. Xie, A. Agrawal, W. keng Liao, and A. N. Choudhary, "Learning to group web text incorporating prior information." in *ICDM Workshops*. IEEE, 2011, pp. 212–219.

[9] K. M. A. Chai, H. L. Chieu, and H. T. Ng, "Bayesian online classifiers for text classification and filtering," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '02. ACM, 2002, pp. 97–104.

[10] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[11] N. Friedman and M. Goldszmidt, "Learning bayesian networks with local structure." in *Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, 1996, pp. 252–262.

[12] R. P. Tina and S. S. Sherekar, "Performance analysis of nave bayes and j48 classification algorithm for data classification," *Inter. Jour. of Computer Science and Applications*, pp. 256–261, 2013.

[13] Sentiwordnet. [Online]. Available: http://sentiwordnet.isti.cnr.it/

[14] Restfb. [Online]. Available: http://restfb.com/

[15] Weka 3: Data mining software in java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[16] Arff. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/arff.html

[17] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.