

# Leveraging Emotion Features in News Recommendations

Nastaran Babanejad  
York University  
Toronto, Canada  
nasba@cse.yorku.ca

Ameeta Agrawal  
York University  
Toronto, Canada  
ameeta@cse.yorku.ca

Heidar Davoudi  
Ontario Tech University  
Oshawa, Canada  
heidar.davoudi@uoit.ca

Aijun An  
York University  
Toronto, Canada  
aan@cse.yorku.ca

Manos Papagelis  
York University  
Toronto, Canada  
papaggel@eecs.yorku.ca

## ABSTRACT

Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. As a specific application domain, news recommender systems aim to give the most relevant news article recommendations to users according to their personal interests and preferences. Recently, a family of models has emerged that aims to improve recommendations by adapting to the contextual situation of users. These models provide the premise of being more accurate as they are tailored to satisfy the continuously changing needs of users. However, little attention has been paid to the emotional context and its potential on improving the accuracy of news recommendations. The main objective of this paper is to investigate whether, how and to what extent *emotion features* can improve recommendations. Towards that end, we derive a large number of emotion features that can be attributed to both items and users in the domain of news. Then, we devise state-of-the-art *emotion-aware recommendation models* by systematically leveraging these features. We conducted a thorough experimental evaluation on a real dataset coming from news domain. Our results demonstrate that the proposed models outperform state-of-the-art non-emotion-based recommendation models. Our study provides evidence of the usefulness of the emotion features at large, as well as the feasibility of our approach on incorporating them to existing models to improve recommendations.

## CCS CONCEPTS

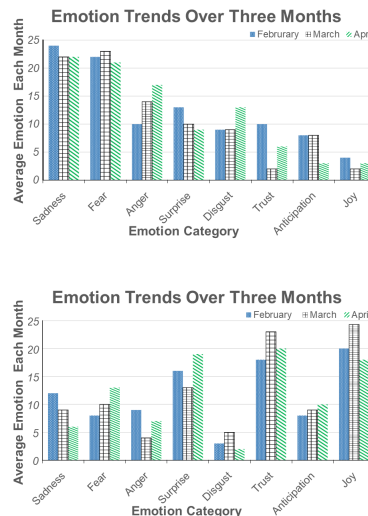
• **Information systems** → **Recommender systems**; **Sentiment analysis**; • **Computing methodologies** → **Neural networks**;

## KEYWORDS

news recommender systems, contextual information, emotion features

## 1 INTRODUCTION

Recommender systems (RS) have widely and successfully been employed in domains as diverse as news and media, entertainment, e-commerce and financial services, to name a few. The main utility of such systems is their ability to suggest items to users that



**Figure 1: Illustrative example of emotions expressed in articles read by two different users,  $U_1$  (top) and  $U_2$  (bottom), over a three month period. Can we leverage the emotional context to improve recommendations?**

they might like or find useful. Traditionally, research on recommendation algorithms has focused on improving the accuracy of predictive models based on a combination of descriptive features of the items and users themselves (e.g., user behavior, interests and preferences) and the history of a user's interactions with the items through ratings, reviews, clicks and more [19]. However, little attention has been paid to the *emotional context* and its relation to recommendations.

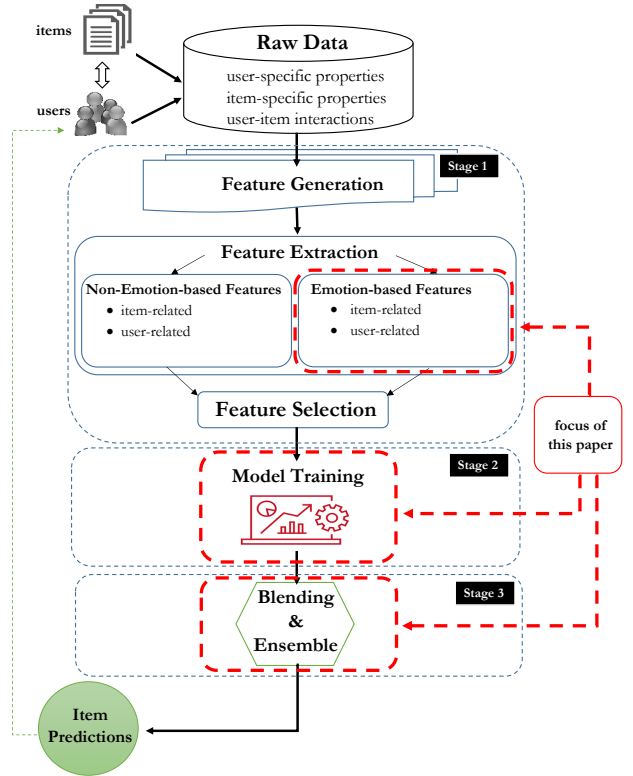
While emotions can be manifested in various ways, we focus on emotions expressed in *textual information* that is associated with *items* or *users* in the system. For example, the content of a news article, the content of an online review or the lyrics of a song are good examples of textual information *directly* associated with an item's emotional context. On the other hand, the *emotional profile* of a user can be determined through *explicit* or *implicit* feedback of users to items. Explicit feedback, such as providing ratings and/or submitting reviews to items, can represent an accurate reflection of a user's opinion about the item, but it is considered an intrusive process that disrupts the user-system interaction and negatively

impacts user experience [31]. In addition, while it might be available for certain domains (e.g. product recommendations [7], movie recommendations [28], etc.), it is not easily obtainable in domains such as news, where users typically interact with items at a fast pace and are less inclined to provide feedback. In the absence, sparsity or high cost of acquisition of explicit feedback, incorporating implicit feedback, which is generally abundant and non-intrusive, might be beneficial. Therefore, we focus on *indirectly* capturing the emotional context of users’ activity by monitoring their interactions with items over time. For instance, one can monitor the tone of the stories in news article users are reading. Effectively, this information can be used to model a user’s historical or temporal emotional profile.

To further motivate this, consider Figure 1 that illustrates the emotional profiles of two users,  $U_1$  and  $U_2$ , based on eight basic emotions, expressed in articles read by them over a period of three months. One can notice that emotions of *sadness* and *fear* are mostly expressed in the articles read by  $U_1$  while other emotions, such as *joy* are less expressed. In addition, one can observe trends such as the expression of *anger* increasing over time. On the other hand, for  $U_2$ , the emotions of *joy* and *trust* are mostly expressed and other emotions, such as *disgust* are less expressed. Moreover, emotions of *fear* and *anticipation* are increasingly expressed in the articles read by this user. Although, the emotional tone derived from news articles read by a user cannot justify the personality and state of mind of the user, it can be considered as the taste or preference of the user, where it shows the type of articles they are more interested in. Inspired by these observations, recent advancement in methods for emotion detection and the success of emotion-aware recommendation algorithms, the main motivation of our research is to investigate *whether, how and to what extent* emotion features can improve the accuracy of recommendations.

**The Problem.** More formally, the recommendation task can be described as follows. Let a set of  $m$  users  $\mathcal{U} = \{u_1, u_2, \dots, u_m\}$  and a set of  $n$  items  $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ . Let us also assume that each user  $u_i$  has already interacted with a set of items  $\mathcal{I}_{u_i} \subseteq \mathcal{I}$  (e.g., consumed news articles). Then, the problem is to accurately predict the probability  $p_{u_a, i_j}$  with which a user  $u_a \in \mathcal{U}$  will like item  $i_j \in \mathcal{I} \setminus \mathcal{I}_{u_a}$ . The task can also take the form of recommending a set  $\mathcal{I}_k \subseteq \mathcal{I} \setminus \mathcal{I}_{u_a}$  of  $k$  items that the user will find most interesting (top- $k$  recommendations). For example, in the news domain, the task is that of recommending an unread article.

**Challenges & Approach.** In order to evaluate the importance of the emotional context to recommendations, we had to incorporate emotional features [33, 42] to state-of-the-art recommendation algorithms and evaluate their accuracy performance. Figure 2, shows a schematic diagram of the emotion-aware recommendation algorithm process we designed, which consists of three main stages: i) *feature engineering*, ii) *model training*, and iii) *blending & ensemble learning*. Each of these components, define a number of challenges that need to be addressed. During *feature engineering*, we had to generate a number of features attributed to both users and items. Emphasis was given in capturing the most important non-emotional and emotional features for the prediction task. Once features are extracted, off-the-shelf *feature selection* methods are employed to select a subset of them that are more relevant for use in model



**Figure 2: Overview of an emotion-aware recommendation system and the focus of the main contributions of the paper.**

construction. During *model training*, we experiment with a number of state-of-the-art models for generating recommendations. During *blending & ensemble* we combine alternative models to obtain better predictive models than any of the constituent models alone.

**Contributions.** The major contributions of this paper are as follows:

- We systematically identify, extract and select the most relevant emotion-based features for use in news recommendation models. These features are associated with both items (e.g., news articles) and users (e.g., readers).
- We devise a number of state-of-the-art models for generating recommendations that incorporate the additional emotion features. These models include variations of *gradient boosting decision trees*, *deep matrix factorization methods* and *deep neural network architectures*. In addition, we use ensembling methods to increase the predictive performance by blending or combining the predictions of multiple constituent models.
- We propose EMOREC, an emotion-aware recommendation model, which demonstrates the best accuracy performance in news recommendation task. EMOREC itself is an ensemble model.
- We conduct a thorough experimental evaluation on a real dataset coming from news domain. Our results demonstrate that the emotion-aware recommendation models consistently outperform state-of-the-art non-emotion-based recommendation models. Our study provides evidence of the

usefulness of the emotion features at large, as well as the feasibility of our approach on incorporating them to existing models to improve recommendations.

## 2 RELATED WORK

Prior research has found a range of features to be useful in the context of news recommender systems, such as user location [14], time of the day [25], demographic information [20], or article social media profile [47]. However, emotion, which is one of the important elements of human nature that has a significant impact on our behavior and choices [46], has received little attention. A number of studies in the area of psychology, neurology, and behavioral sciences have shown that individuals' choices are related to their feelings and mental moods [23].

In the context of recommender systems, one of the earliest works, Gonzalez et al. [16], pointed out that emotions are crucial for users' decision making and that users transmit their decisions together with emotions. Tkalcic et al. [39] introduced a unifying framework for using emotions in user interactions with a recommender system, and suggested that while an implicit approach of user feedback may be less accurate, it is well suited for user interaction purposes since the user is not aware of it [38].

While emotions as features have been studied in movie recommendations [27, 28], music recommendations [17] and restaurant recommendations [41], to name a few, much less work has explored the role of emotion features in news recommender systems.

Emotion in news articles has been studied for categorizing news stories into eight emotion categories [2]. Specifically for recommender systems, Parizi and Kazemifard [32] introduced a model for Persian news utilizing both, the emotion of news as well as user's preference. More recently, Mizgajski and Morzy [22] introduced a recommender system for recommending news items by leveraging a multi-dimensional model of emotions, where emotion is derived through user's self-assessed reactions (i.e., explicit feedback) which can be considered as intrusive collection. In contrast to previous studies, our work focuses on studying the role of emotion features in news recommender systems using implicit user feedback.

## 3 FEATURES FOR RECOMMENDATION

This section describes the feature extraction procedure which is utilized in our proposed framework. The features are grouped into two main categories: (i) *emotion-based features* for items and users, and (ii) *non-emotion-based features* for items and users.

### 3.1 Emotion-based Features

The main objective of this paper is to improve the performance of recommender system by leveraging the user/item emotion features. Figure 3 shows an example of textual content of items (i.e., an article) in news domains. As it can be observed, there are several words such as *win* and *gratifying*, expressing the emotion of *happiness*. Moreover, interjections such as *yay* and *oh* can be indicators of different emotions [15]. In this section, we describe how we extract such features to improve the recommendation system effectiveness. In order to maintain consistency, each news article is preprocessed by tokenizing into words, removing the stopwords and POS-tagging

#### excerpt of news article

"It was **[gratifying]** **yay**... Who wouldn't want it to keep going?" Walker told NBA.com. "**Oh**, I know teams will be gearing up on me and double-teaming me. But I just want to **win**, man. I want to get back to the playoffs any way possible. I don't care what I average the rest of the year."

Figure 3: Example emotions expressed in textual content

Table 1: Emotion Resources

| Resources            | Size           | Emotion Taxonomy |
|----------------------|----------------|------------------|
| WordNet-Affect [36]  | 4787 words     | Several          |
| ISEAR [43]           | 7600 sentences | ISEAR            |
| NRC [24]             | 14,182 words   | Plutchik         |
| SentiWordNet 3.0 [3] | 11,000+ synset | Sentiments       |

to extract nouns, verbs, adverbs and adjectives. In particular, we focus on two approaches for computing emotion features: sentiment analysis, which classifies text into *neutral*, *positive* and *negative* sentiments, and emotion analysis which categorizes text into emotions such as *happiness*, *sadness*, *anger*, *disgust*, *fear* and so on. Note that we extract emotion features for both users and items.

#### 3.1.1 Item Emotion-based Features.

**Number of Emotion Words:** This feature represents the number of words in an emotion lexicon (i.e., WordNet-Affect, see Table 1) that occur in the item (i.e., news article) more than once.

**Ekman's Emotion Label:** We count the number of emotion words occurring in the text document for each emotion type (Ekman's six emotion categories [12]) and then the text is assigned an emotion label with the highest number of emotion words appearing in the text. If more than one emotion category has the highest count, 0 is assigned to this feature, leaving the next feature to indicate mixed emotions. A combination of different lexicons (WordNet-Affect and NRC, see Table 1) is used to find the emotion labels. We use multiple resources to have a bigger set of emotion words for each emotion.

**Mixed Emotions:** This feature indicates whether an item has more than one document-level emotion labels based on Ekman's emotion model (i.e., if two or more emotions have the highest score, this feature is valued at 1, otherwise 0). Since the initial annotation effort (previous feature) illustrated that in many cases, a sentence can exhibit more than one emotion, we have an additional category called mixed emotion to account for all such instances.

**Sentiment Feature:** The text is classified into three categories: *positive*, *negative* and *neutral*. We utilize the approach introduced in [29] and use SentiWordNet [3].

**Interjections:** This feature counts the number of interjections in a document. A short sound, word or phrase spoken suddenly to express an emotion, e.g., *oh*, *look out!*, *ah*, are called interjections<sup>1</sup>. Our preliminary analysis found that interjections were common

<sup>1</sup>List of interjections derived from: i) <https://surveyanyplace.com/the-ultimate-interjection-list>, ii) <https://7esl.com/interjections-exclamations>, and iii) <https://www.thoughtco.com/interjections-in-english-1692798>

in quotes in news articles, which can be detected for potential emotions.

**Capitalized Words:** This feature counts the number of words in a document with all uppercase characters. People use capital words to express an emotion [40] and make it bold to the readers (e.g., *I said I am FINE*).

**Punctuation:** Two features are included to model the occurrence of question marks and exclamation marks repeated more than two times in a document. Using punctuation can clarify the emotional content of the texts that are sometimes easy to miss [40].

**Grammatical Markers and Extended Words:** This feature counts the number of times words with a character repeated more than two times (e.g., haaappy or oh yeah!!????) [6] as excessive use of letters in a word (e.g., repetition) is one way to emphasize feelings.

**Plutchik Emotion Scores:** First, we measure the semantic relatedness score between a word  $W_i$  in the text and an emotion category  $C_j$  in the NRC lexicon (see Table 1) as follows [1]:

$$PMI(W_i, C_j) = \sqrt[n]{\prod_{k=1}^n PMI(W_i, C_j^k)} \quad (1)$$

where  $C_j^k$  ( $k = 1 \dots n$ ) is the  $k^{\text{th}}$  word of emotion category  $C_j$ .  $PMI$  is the Pointwise Mutual Information calculated as follows:

$$PMI(W_i, C_j^k) = \log \frac{P(W_i, C_j^k)}{P(W_i)P(C_j^k)} \quad (2)$$

where  $P(W_i)$  and  $P(C_j^k)$  are the probabilities that  $W_i$  and  $C_j^k$  occur in a text corpus, respectively, and  $P(W_i, C_j^k)$  is the probability that  $W_i$  and  $C_j^k$  co-occur within a sliding window in the corpus. Finally, we calculate the average, maximum and minimum of score for all words in the text for each emotion category and consider each as an individual feature.

### 3.1.2 User Emotion-based Features.

As we do not have access to users' explicit emotion towards items, we develop users' *implicit* emotional profile based on their historical interactions with items. By computing the emotion profile of the items with which a user is interacting, we derive the emotional taste of the user over that period of time over the set of items.

**User Emotions Across Items:** We determine the emotion score (i.e., Plutchik's emotion scores) for the last accessed item before subscription as well as for the last 20 items accessed by the user. Then, we pick the top 3 frequent emotions.

**User Emotions Across Categories:** We determine the emotion of categories of items (e.g., sports in news domain) accessed by a user by counting the number of items assigned to an emotion in a specific category, with the most frequent emotion considered as the emotion of the category. The feature is calculated for the whole history of the user.

## 3.2 Non-Emotion-based Features

Non-emotion-based features can also be classified into item-based and user-based features.

### 3.2.1 Item Non-Emotion-based Features.

**Item Topic:** We extract topics in the article using Latent Dirichlet Allocation (LDA) [5]. In LDA, each topic is a distribution over

words, and each document is a mixture of topics. The number of topics for the news articles are 112, which were chosen empirically to minimize the perplexity score of the LDA result. Thus, the item topic is represented by a vector of length 112.

**Topic Label:** We use `lda2vec` [26] to generate and label the topics in an item (i.e., document), where each generated topic is labeled by one of its top  $k$  words which is most semantically similar to the other words in the top  $k$  word list. We then label the item (i.e., document) with the label of the most coherent topic among the top  $m$  topics of the document. The word vector of this label word is used as the value for this feature.

**TF-IDF:** This feature represents items as n-grams (unigram, bigram, trigram) with the TF-IDF weighting approach [21].

**Coherence:** We first calculate the cosine similarity scores between all pair of words in an item using `word2vec` pre-trained word vectors<sup>2</sup>, and then record average of similarity scores, standard deviation of similarity scores, the lowest score that is higher than the standard deviation, and the highest score that is lower than the standard deviation as four features.

**Potential to Trigger Subscription:** This feature represents the total number of times the item was requested right before a paywall was presented to a user who subsequently made a subscription [9, 10]. In a subscription-based item delivery model a paywall is the page asking for subscription before allowing an unsubscribed user to continue accessing items.

### 3.2.2 User Non-emotion-based Features.

**Visit Count:** We calculate the average number of items (articles) accessed by a user per visit. A visit is terminated if a user is inactive for more than 30 minutes.

**User Spent Time:** Two features are represented. One is the average time the user spent per item, and the other is the average time the user spent per visit.

**User Interest in Subcategory:** This feature represents the empirical probability of subcategory  $s$  given a user  $u$  and a category  $c$  denoted as  $P(s|u, c)$ . For example,  $P(election|u, politics)$  can be determined by the total number of articles the user read on election over the total number of articles that the user read on politics. In our experiments, the categories and subcategories were provided with the dataset and we consider only the top 50 most frequently visited subcategories for this feature.

**User Latent Vector:** We calculate the latent vector for each user based on matrix factorization introduced in [37]. This feature is chosen so that we can compare our method with the Deep Matrix Factorization model in [44], a state-of-the-art recommendation method, which uses this feature as input for a deep neural network.

## 3.3 Feature Selection

One of the critical steps after feature extraction is to select important features for recommendation. Table 2 reports the most important features according to gain importance score for the news data set. We evaluate feature importance by averaging over 10 training runs of a gradient boosting machine learning model XGBoost [8] to reduce variance<sup>3</sup>. Also, the model is trained using early stopping

<sup>2</sup><https://code.google.com/archive/p/word2vec/>

<sup>3</sup>Variance refers to the sensitivity of the learning algorithm to the specifics of the training data (e.g., the noise and specific observations).

**Table 2: List of Emotion/Non-emotion Feature Importance**

| Emotion Features                       | Gain Score |
|--|------------|
| Plutchik emotion scores                | 3200.86    |
| User emotions across items             | 1985.36    |
| User emotions across categories        | 1850.33    |
| Ekman’s emotion label                  | 1101.38    |
| Punctuation                            | 910.55     |
| Grammatical markers and extended words | 860.13     |
| Interjections                          | 773.12     |
| Capitalized words                      | 640.21     |
| Mixed emotions                         | 526.97     |
| Sentiment features                     | 360.68     |
| Non-emotion Features                   | Gain Score |
| User latent vector                     | 3640.87    |
| Potential to trigger subscription      | 2974.46    |
| User interest in subcategory           | 1530.28    |
| Topic labeling                         | 1421.19    |
| User spent time                        | 1110.57    |
| Visit count                            | 920.53     |
| Item topic                             | 867.12     |
| Coherence                              | 685.23     |
| TF-IDF                                 | 410.29     |

with a validation set to prevent over-fitting to the training data. By using the zero importance function, we find features that have zero importance according to XGBoost.

## 4 RECOMMENDATION MODEL

In this section, we introduce a tailored structure of an Emotion-aware Recommender System Model (EMOREC) for personalized recommendation. Our final model is an ensemble model of three models leveraging both emotion/non-emotion-based features. We describe the structure of the proposed model and the training methods next.

### 4.1 Model Training

**Model 1 (Boost Model):** Gradient Boosting Decision Tree (GBDT) methods are among the most powerful machine learning approaches which have been effectively used in many domains [13] including recommendation[45]. The basic idea in GBDT approaches is to learn a set of *base/weak learners* (i.e., decision trees) sequentially by using different training splits. More precisely, at each step, we learn a new base model by fitting it to the error residuals (i.e., difference between the current model predictions and the actual target values) at that step. The new model outcome is the previous model outcome plus the (weighted) new base learner outcome. Eventually, the final model outcome is the weighted average of all base learners outcome, where the weights are learned jointly with the base learners. We train two state-of-the-art GBDT models, namely, XGBoost [8] and Catboost [11], on our training datasets with the features selected in Section 3.3 as the input.

XGBoost uses pre-sorted/histogram-based algorithm to compute the best split while CatBoost uses *ordered boosting*, a permutation based algorithm, to learn the weak learners effectively. Moreover,

XGBoost uses one-hot encoding before supplying categorical data, but CatBoost handles categorical features directly. We train both models individually (three base models for each). The final model output (i.e., probability that a user is interested in an item) is the combination of all base models outcomes:

$$\sum_i^6 \alpha_i p_i \quad (3)$$

where  $p_i$  is the probability that the user is interested in the item according to base model  $i$  and  $\alpha_i$  is the weight of base model  $i$  learned by XGboost/Catboost.

**Model 2 (Deep Neural Network (DNN)):** Figure 4 shows our proposed Deep Neural Network architecture for leveraging the emotion features (and other commonly available features) for the recommendation purpose. The input is divided into four groups[4]: i) user non-emotion based features, ii) item non-emotion based features, iii) user emotion-based features, and iv) item emotion-based features. For the categorical inputs, we utilize one-hot encoding (the second layer is look-up embeddings mapping each categorical feature to a fixed length embedding vector). In the architect “Dense Layer” can be formalized as:  $Dense(x) = f(Wx + bias)$  where  $W$  and  $bias$  are parameters,  $x$  is the layer input and  $f$  is the activation function (for linear layer  $f$  is the identity function). We use  $L_2$  regularization to prevent over-fitting in embedding layer and use back-propagation to learn the parameters.

**Model 3 (Deep Matrix Factorization (Deep MF)):** Inspired by the models proposed in [18, 44], we built our Deep MF (Figure 5) to leverage extra user/item features (i.e., emotion and non-emotion features) in the recommendation prediction task. In [44], they construct a user-item matrix with explicit ratings and implicit preference feedback, then with this matrix as the input, they present a deep neural architecture to learn a low dimensional space for the representation of both users and items. In [18], by replacing the inner product with a neural architecture, they learn an arbitrary function to capture the interactions between user and item latent vectors. Different from their work, we focused on modeling the user/item with rich extra features, such as non-emotion and emotion based features, as well as using embedding vectors learned in our DNN model. The input of our proposed model is the same as the DNN model where the categorical features are encoded using one hot vectors. The second layer is the look-up embedding. In this layer, we have both MF embedding vectors, which we estimate through the learning process, and DNN embedding vectors, which are concatenation of embedding vectors (for each similar input group) learned from DNN model (they are fixed in this model). Generalized Matrix Factorization (GMF) layer combines two embeddings using dot product and applies some non-linearity. Similar to DNN model, the output of the model is the probability that a user is interested in an item.

**Ensemble/Blending Model:** The final model EMOREC was the weighted average of the three models’ predictions. We use Nelder-Mead Method [30] to find the optimum weights of each models.

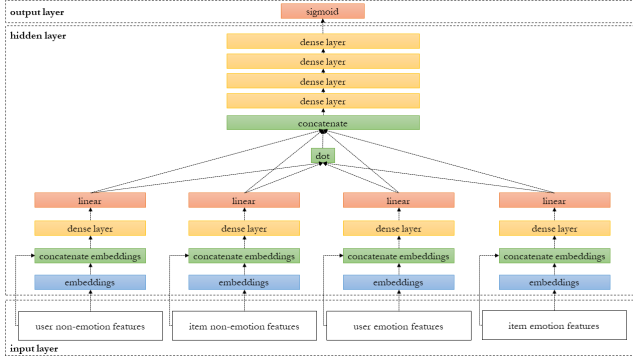


Figure 4: The Structure of Our DNN Model

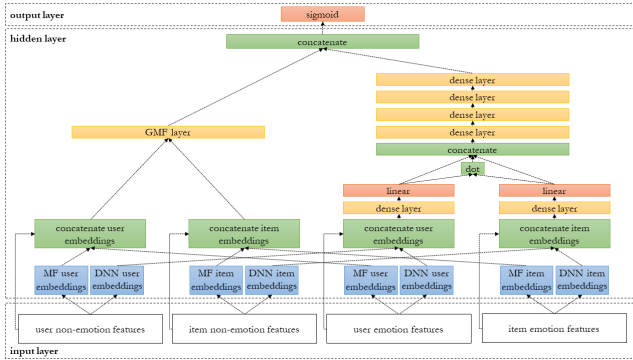


Figure 5: The Structure of Our Deep MF Model

## 5 EXPERIMENTS

In this section, we introduce the data, evaluation protocols and the specific configurations used in our experiments.

### 5.1 Data

Our experiments are conducted on a real-world news dataset. The Globe and Mail is one of the major newspapers<sup>4</sup> in Canada. We use the data spanning from January to July 2014 (a 6-month period) in our experiments where the data in the first 4 months were used for training, and the last 2 months for testing. The dataset contains information for 359,145 articles in total and 88,648 users in total, out of which 17,009 became subscribers during this period, and 71,639 were non-subscribers. Every time a user reads an article, watches a video or generally takes an action in the news portal, the interaction is recorded as a hit. Typically, a hit contains information like date, time, user id, visited article, special events of interest like subscription, sign in, and so on.

### 5.2 Evaluation Metrics

We use F-score to measure the predictive performance of a recommender system. For each user in the test data set, we use the original set of read articles in the test period as the ground truth, denoted as  $T_g$ . Assuming the set of recommended news articles

Table 3: Results of our Models on News Dataset (F-score)

| Model   | Non-Emo | All          |
|---|---------|--------------|
| Single Boost Model                            | 70.19   | 70.86        |
| Boost Blend                                   | 70.69   | 71.50        |
| Deep MF                                       | 72.93   | 73.29        |
| Single DNN Model                              | 70.88   | 73.00        |
| DNN Ensemble                                  | 73.62   | 74.30        |
| Boost Blend + Deep MF                         | 73.07   | 74.98        |
| Boost Blend + DNN Ensemble                    | 74.00   | 74.23        |
| Deep MF + DNN Ensemble                        | 74.61   | 75.10        |
| EMOREC (Boost Blend + Deep MF + DNN Ensemble) | 78.20   | <b>80.30</b> |

for the user is  $T_r$ , precision, recall, and F-measure are defined as follows:

$$Precision = \frac{|T_g \cap T_r|}{|T_r|}, \quad Recall = \frac{|T_g \cap T_r|}{|T_g|}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F-score on a test data set is the average over all the users in the test data set.

### 5.3 Comparing Recommendation Models with and without Emotion Features

Our main objective is to see whether the use of emotion features will boost the performance of recommendation models. For such a purpose we run the three state-of-the-art recommendation models described in the last section and some ensembles formed by these models with and without emotion features. The models used in our evaluation are as follows:

- *Single Boost Model*: We run XGBoost and Catboost separately to make predictions and collect the average of their F-scores.
- *Boost Blend*: This is the 6-model ensemble described in Model 1 in Section 4.1.
- *Deep MF*: This is the deep matrix factorization model described in Section 4.1.
- *Single DNN model*: We run the DNN model for 5 times with the same hyperparameters but different random seeds and collect the average result over 5 runs.
- *DNN Ensemble*: An ensemble of 5 DNN models with different hyperparameters (e.g., different learning rates, etc.) is run 5 times each with a different random seed. The average result over the 5 runs is collected.
- *Boost Blend + Deep MF*: This is an ensemble consisting of Boost Blend and Deep MF.
- *Boost Blend + DNN Ensemble*: This an ensemble consisting of Boost Blend and DNN Ensemble.
- *Deep MF + DNN Ensemble*: This is an ensemble consisting of Deep MF and DNN Ensemble.
- *Boost Blend + Deep MF + DNN Ensemble*: an ensemble consisting of Boost Blend, Deep MF and DNN Ensemble.

We train each of the above models using the training data of our data set and use the trained model to make recommendations by predicting a user's interest in an item in the test data. Table 3 shows

<sup>4</sup><https://www.theglobeandmail.com/>

**Table 4: Comparison of EMOREC with State-of-the-art Baselines on News Dataset (F-score)**

| Model                                   | Non-Emo | All          |
|---|---------|--------------|
| Basic MF                                | 69.10   | 71.23        |
| FDEN and GBDT                           | 72.02   | 73.28        |
| Truncated SVD-based Feature Engineering | 73.12   | 74.01        |
| EMOREC                                  | 78.20   | <b>80.30</b> |

the results (in F-score) of using these recommendation methods with and without emotion features on the news data set, where the whole set of emotion features described in Section 3.3 is used in the results for "All", while none of the emotion features is used in the results for "Non-Emo". As can be seen, adding emotion features improves the predictive performance for all the recommendation methods. Among the single recommendation models (i.e., Single Boost Model, Deep MF and Single DNN Model), Deep MF performs the best. The results also show that ensemble methods perform better than single/component models. The best performance is produced by the largest ensemble (i.e., *Boost Blend + Deep MF + DNN Ensemble*). We refer to this best-performing model as our EMOREC model.

#### 5.4 Comparison with Other Baselines

We also compare our EMOREC model with the following three state-of-the-art recommendation methods with well-tuned parameters (that is, the parameters are optimally tuned to ensure the fair comparison). The objective is to investigate whether emotion features can smarten up these recommender systems. A brief description of these three models is as follows:

*Basic MF*: This is the simple matrix factorization model where used for discovering latent features between two entities (i.e., user and articles)[37]. Both user preferences and item characteristics are mapped to latent factor vectors. Each element of the item-specific factor vector measures the extent to which the item possesses one feature. Accordingly, each element of the user-specific factor vector measures the extent of the user preferences in that feature.

*FDEN and GBDT*: an ensemble of different models, including Field-aware Deep Embedding Networks and Gradient Boosting Decision Trees [4]. The predictions of FDENs are from a bagging ensemble using the arithmetic mean of many networks, each of which has slight differences on hyper-parameters, including the forms of the activation.

*Truncated SVD-based Feature Engineering*: a gradient boosted decision trees model with truncated SVD-based embedding features [34]. To overcome the cold start problem, a truncated SVD-based embedding features were created using the embedding features with four different statistical based features (users, items, artists and time), the final model was the weighted average of the five models' predictions.

The results are illustrated in Table 4, which shows that emotion features can also improve the recommendation performance of these three state-of-the-art baselines. In addition, our EMOREC model performs significantly better than these three baselines in

**Table 5: Effect of Individual Emotion Features (F-score)**

| Emotion Features                         | News  |
|--|-------|
| ALL emotion features                     | 80.30 |
| - Sentiment features                     | 78.15 |
| - Mixed emotions                         | 76.90 |
| - Capitalized words                      | 76.21 |
| - Interjections                          | 75.84 |
| - Grammatical markers and extended words | 75.23 |
| - Ekman's emotion label                  | 74.98 |
| - Punctuation                            | 75.17 |
| - User emotions across categories        | 74.15 |
| - User emotions across items             | 73.23 |
| - Plutchik emotion scores                | 72.10 |

**Table 6: Effect of Top Three Emotion Features (*Plutchik emotions, User emotions across categories, and User emotions across items*) on State-of-the-art Models**

| Model               | No Emotion | Top Three Emotion |
|---------------------|------------|-------------------|
| Basic MF            | 69.10      | 70.38             |
| Boost Blend         | 70.69      | 71.00             |
| FDEN and GBDT       | 72.02      | 72.77             |
| Deep MF             | 72.93      | 73.01             |
| Truncated SVD-based | 73.12      | 73.60             |
| DNN Ensemble        | 73.62      | 73.98             |

both cases of using emotion features and not using emotion features.

#### 5.5 Effect of Individual Emotion Features

Table 5 presents the results of a feature ablation study in order to further understand the effect of individual emotion features used in EMOREC. In each run of this study, we keep all the features except one type of emotion features. The results indicate that removing Plutchik emotion scores (item feature), User emotions across categories and User emotions across items (user features) lead to considerable decline in the performance. It also shows that our model is able to capture useful implicit user emotion effectively.

To further validate the effectiveness of the top emotion features as learned from our experiments, we run a further experiment incorporating only the top three emotion features (i.e., Plutchik emotions, User emotions across categories, and User emotions across items) on six state-of-the-art recommendation models. As the results in Table 6 show, only using these three emotion features can also improve the recommender systems, with Basic MF showing the most gain.

## 6 CONCLUSIONS

Motivated by the recent development in emotion detection methods (in textual information), we considered the problem of leveraging emotion features to improve recommendations. Towards that end, we derived a large number of emotion features that can be attributed to both items and users in news domain and can provide an

emotional context. Then, we devised state-of-the-art *non-emotion* and *emotion-aware recommendation models* to investigate *whether, how and to what extent* emotion features can improve recommendations. To the best of our knowledge, this is the first attempt to systematically and broadly evaluate the utility of a number of emotion features for the recommendation task. Our results indicate that emotion-aware recommendation models consistently outperform state-of-the-art non-emotion-based recommendation models. Furthermore, our study provided evidence of the usefulness of the emotion features at large, as well as the feasibility of our approach on incorporating them to existing models to improve recommendations.

As a more tangible outcome of the study, we proposed EMOREC, an emotion-aware recommendation model, which demonstrates the best predictive performance in news recommendation task. EMOREC itself is an ensemble model combining three models (*Boost Blend + Deep MF + DNN Ensemble*). It significantly outperforms other state-of-the-art recommendation methods evaluated in our experiments. We also evaluated the proposed emotion features individually. Among the emotion features examined, the Plutchik emotion scores of items (obtained by computing PMI scores between words) and user emotion profiles (based on the emotion scores of the items that the user accessed) are the most important.

Employing emotional context in recommendations appears to be a promising direction of research. While the scope of our current study is limited to emotions extracted by textual information, there is evidence that emotions can be extracted through other means of communication, such as *audio* and *video*, or other cues [35].

## 7 ACKNOWLEDGEMENTS

This work is funded by Natural Sciences and Engineering Research Council of Canada (NSERC), The Globe and Mail, and the Big Data Research, Analytics, and Information Network (BRAIN) Alliance established by the Ontario Research Fund Research Excellence Program (ORF-RE). We would like to thank The Globe and Mail for providing the dataset used in this research. In particular, we thank Gordon Edall of The Globe and Mail for their insights and collaboration in our joint project.

## REFERENCES

- [1] Ameeta Agrawal and Aijun An. 2012. Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, Macau, China, 346–353. <https://doi.org/10.1109/WI-IAT.2012.170>
- [2] Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Emotion Sensitive News Agent (ESNA): A system for user centric emotion sensing from the news. *Web Intelligence and Agent Systems* 8, 4 (2010), 377–396.
- [3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, Vol. 10. 2200–2204.
- [4] Bing Bai and Yushun Fan. 2017. Incorporating Field-aware Deep Embedding Networks and Gradient Boosting Decision Trees for Music Recommendation. In *The 11th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, London, England, 7.
- [5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (March 2003), 993–1022.
- [6] Mondher Bouazizi and Tomoaki Otsuki. 2016. A Pattern-Based Approach for Sarcasm Detection on Twitter. *IEEE Access* 4 (2016), 5477–5488. <https://doi.org/10.1109/ACCESS.2016.2594194>
- [7] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender Systems Based on User Reviews: The State of the Art. *User Modeling and User-Adapted Interaction* 25, 2 (June 2015), 99–154. <https://doi.org/10.1007/s11257-015-9155-5>
- [8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [9] Heidar Davoudi, Aijun An, Morteza Zihayat, and Gordon Edall. 2018. Adaptive Paywall Mechanism for Digital News Media. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, New York, NY, USA, 205–214. <https://doi.org/10.1145/3219819.3219892>
- [10] H. Davoudi, M. Zihayat, and A. An. 2017. Time-Aware Subscription Prediction Model for User Acquisition in Digital News Media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Houston, Texas, USA, 135–143. <https://doi.org/10.1137/1.9781611974973.16>
- [11] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. (Oct. 2018).
- [12] Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion* 3 (1984), 19–344.
- [13] Ji Feng, Yang Yu, and Zhi-Hua Zhou. 2018. Multi-Layered Gradient Boosting Decision Trees. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 3551–3561.
- [14] Blaž Fortuna, Carolina Fortuna, and Dunja Mladenčić. 2010. Real-time news recommender system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 583–586.
- [15] Cliff Goddard. 2014. Interjections and Emotion (with Special Reference to ÅÄJJSurpriseÅÄ and ÅÄJJDismustÅÄ). *Emotion Review* 6, 1 (Jan. 2014), 53–63. <https://doi.org/10.1177/1754073913491843>
- [16] Gustavo Gonzalez, Josep Lluís de la Rosa, Miquel Montaner, and Sonia Delfin. 2007. Embedding Emotional Context in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW '07)*. IEEE Computer Society, Washington, DC, USA, 845–852.
- [17] Byeong-Jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. 2010. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications* 47, 3 (2010), 433–460.
- [18] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, Perth, Australia, 173–182. <https://doi.org/10.1145/3038912.3052569>
- [19] Dhruv Khattar, Vaibhav Kumar, Manish Gupta, and Vasudeva Varma. 2018. Neural Content-Collaborative Filtering for News Recommendation. In *NewsIR'18 Workshop*. NewsIR@ECIR, Grenoble, France, 1395–1399.
- [20] Hong Joo Lee and Sung Joo Park. 2007. MONERS: A news recommender for the mobile web. *Expert Systems with Applications* 32, 1 (2007), 143–150.
- [21] Christopher D Manning, Prabhakar Raghavan, and Hinrich SchÅÄjtze. 2009. Introduction to Information Retrieval. (2009), 569.
- [22] Jan Mizgajski and Mikolaj Morzy. [n. d.]. Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction* ([n. d.]), 1–35.
- [23] Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *In Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*. Association for Computational Linguistics, Copenhagen, Denmark, 34ÅÄ\$49. <https://arxiv.org/abs/1708.03700>
- [24] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a WordÅÄ\$Emotion Association Lexicon. *Computational Intelligence* 29, 3 (Aug. 2013), 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- [25] Alejandro Montes-García, Jose María Álvarez-Rodríguez, Jose Emilio Labra-Gayo, and Marcos Martínez-Merino. 2013. Towards a journalist-based news recommendation system: The Wesomender approach. *Expert Systems with Applications* 40, 17 (2013), 6735–6741.
- [26] Christopher E. Moody. 2016. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *arXiv:1605.02019 [cs]* (May 2016). <http://arxiv.org/abs/1605.02019> arXiv: 1605.02019.
- [27] Yashar Moshfeghi, Benjamin Piwowarski, and Joemon M Jose. 2011. Handling data sparsity in collaborative filtering using emotion and semantic based features. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 625–634.
- [28] Ante Odić, Marko Tkalčić, Jurij F Tasić, and Andrej Košir. 2013. Predicting and detecting the relevant contextual information in a movie-recommender system. *Interacting with Computers* 25, 1 (2013), 74–90.
- [29] Sylvester Olubolu Orimaye, Saadat M. Alhashmi, and Siew Eu-gene. 2012. Sentiment Analysis Amidst Ambiguities in Youtube Comments on Yoruba Language (Nollywood) Movies. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 583–584. <https://doi.org/10.1145/2187980.2188138>

- [30] Yoshihiko Ozaki, Masaki Yano, and Masaki Onishi. 2017. Effective hyperparameter optimization using Nelder-Mead method in deep learning. *IPSJ Transactions on Computer Vision and Applications* 9, 1 (Nov. 2017), 20. <https://doi.org/10.1186/s41074-017-0030-7>
- [31] Maja Pantic and Alessandro Vinciarelli. 2009. Implicit human-centered tagging [Social Sciences]. *IEEE Signal Processing Magazine* 26, 6 (2009), 173–180.
- [32] Ali Hakimi Parizi and Mohammad Kazemifard. 2015. Emotional news recommender system. In *2015 Sixth International Conference of Cognitive Science (ICCS)*. IEEE, 37–41.
- [33] Mikhail Rumiantsev. 2017. *Music adviser : emotion-driven music recommendation ecosystem*. Ph.D. Dissertation. Department of Mathematical Information Technology Olexiy Khriyenko. <https://jyx.jyu.fi/handle/123456789/53196>
- [34] Nima Shahbazi, Mohamed Chahhou, and Jarek Gryz. 2017. Truncated SVD-based Feature Engineering for Music Recommendation. In *The 11th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, London, England, 7.
- [35] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2016. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2016), 17–28.
- [36] Carlo Strapparava, Alessandro Valitutti, and others. 2004. WordNet Affect: an Affective Extension of WordNet. In *LREC*, Vol. 4. European Language Resources Association (ELRA), Lisbon, Portugal, 1083–1086.
- [37] Gábor Takács, István Pilászy, Botyán Németh, and Domonkos Tikk. 2008. Investigation of various matrix factorization methods for large recommender systems. In *2008 IEEE International Conference on Data Mining Workshops*. IEEE, 553–562.
- [38] Marko Tkalčić, Urban Burnik, Ante Odić, Andrej Košir, and Jurij Tasič. 2012. Emotion-aware recommender systems—a framework and a case study. In *International Conference on ICT Innovations*. Springer, 141–150.
- [39] Marko Tkalčić, Andrej Kosir, Jurij Tasivc, and Matevž Kunaver. 2011. Affective recommender systems: the role of emotions in recommender systems. 9–13.
- [40] Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.. In *Fourth International AAAI Conference on Weblogs and Social Media*. Association for Computational Linguistics, Washington, DC, 107–116.
- [41] Blanca Vargas-Govea, Gabriel González-Serna, and Rafael Ponce-Medellin. 2011. Effects of relevant contextual features in the performance of a restaurant recommender system. *ACM RecSys* 11, 592 (2011), 56.
- [42] Karzan Wakil, Rebwar Bakhtyar, Karwan Ali, and Kozhin Alaadin. 2015. Improving Web Movie Recommender System Based on Emotions. *International Journal of Advanced Computer Science and Applications* 6, 2 (2015), 9. <https://doi.org/10.14569/IJACSA.2015.060232>
- [43] H. G. Wallbott and K. R. Scherer. 1986. How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Social Science Information* 25, 4 (Dec. 1986), 763–795. <https://doi.org/10.1177/053901886025004001>
- [44] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep Matrix Factorization Models for Recommender Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 3203–3209. <https://doi.org/10.24963/ijcai.2017/447>
- [45] Qian Zhao, Yue Shi, and Liangjie Hong. 2017. GB-CENT: Gradient Boosted Categorical Embedding and Numerical Trees. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.
- [46] Yong Zheng, Robin Burke, and Bamshad Mobasher. 2013. The Role of Emotions in Context-aware Recommendation. In *RecSys workshop in conjunction with the 7th ACM conference on Recommender Systems*. RecSys workshop in conjunction with the 7th ACM conference on Recommender Systems, Hong Kong, China., 8.
- [47] Morteza Zihayat, Anteneh Ayanso, Xing Zhao, Heidar Davoudi, and Aijun An. 2019. A utility-based news recommendation system. *Decision Support Systems* 117 (2019), 14–27.