Research Paper 1: Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text

Points from abstract:

- The reviews are categorized into three basic classes i.e. negative, positive, and neutral. This paper focuses on Urdu Roman reviews that are obtained by one of the most famous and accessed e-commerce website of Pakistan–Daraz.pk.
- Vector space model, a.k.a bag of word model is applied for feature extraction which are later passed to Support Vector Machines (SVM) for sentiment classification.
- Experiments are conducted on MATLAB Linux server.

Points from Introduction:

- Products are getting hundreds of reviews online affecting the overall perception of the company and their products.
- We examine the reviews of DarazPK and observe that in Pakistan, customers conventionally post the comments either in English language or Roman Urdu language. Roman Urdu, Urdu being native language of Pakistan, is widely used, making it easy for people to correctly express their feelings.
- Daraz.pk is the most commonly used site therefore, the reviews are being filled and thus data is available.
- Unlike other sites like symbios.pk, homeshopping.pk, shophive.com, ishopping.pk, and 24hours.pk/ don't have an option to review the products.
- Reviews determines and examines the user perception assisting the sellers to increase and enhance their products availability and quality hence, affecting online shopping positively.
- Machine learning and lexicon-based learning has been done at great extent. These can be used in Roman Urdu reviews.
- Roman text is usually very challenging to process. Roman Urdu has no standards and no rules therefore, understanding such language is not so easy. For example, word Mein means I/Me, and can be written in various ways such as, Mn, Me, Main, Ma, Men, M, etc. Non-standard word forms make it difficult to process and understand.

Points from Related work:

1. Amazon reviews: (Online, surveys)
    - Big data was collected from amazon and reviews. Recommendation system was used to check the users reviews priority and qualitative analysis was done to check sentiment analysis on the large data.
    - Bootstrapping method was used to extract adopter information reviews from site
    - Drawn a distribution curve of the products of the collected data. Products were divided into categories like product category, number of products, review of product and mean of product.
    - Compared the models and proved their work by lemma, proof of proposition and proof of corolla.
    - A survey was done using Qualtrics, a questionnaire. Data was displayed in the form of tables. Positive and negative reviews were extracted. Reviews were tested by test classifier.

- Logistic regression and L2 regularization were used as baseline of classifier. Drawn ROC curve, F1 measure was also calculated, also measured precision and recall, unigram, n-gram, and histogram.
- Another online reviews data was extracted, and Three classifiers were used and discussed, 1: Passive Aggressive algorithm 2: Language Modeling, 3: Winnow Classifier.
- Reviews were about electronics, book, beauty and home and were categorized into tables.

2. **Product reviews from Twitter:**
   - Twitter has its own features like emoticons, abbreviation, hashtags etc, this creates less recall for lexicon-based method.
   - Hashtags words like # fail was included in their lexicon.
   - POS tagger was used for comparative sentences like iPhone is better than Samsung
   - F measure, recall and precision was used for evaluation measure.
   - **Aim was to perform replication using webis source code to see if it produces comparable result**

MAIN RELATED WORK IN ROMAN LANGUAGES

3. **Product reviews in native Persian language:**
   - Extracted data was then filtered in WEKA2.
   - Nave Bayes algorithm, decision trees, and KNN used. And in the end result of each algo compared.
   - **nave Bayes provides good result as compared to decision tree because of its simplicity and easy use**
   - First created dictionary of Persian words. They have added Persian words in it which was the difficult task of their work.
   - Data set was collected from **digikala** by using web crawler.dataset.
   - Pre-processing was done using spell checker, lemmatization, pos, stop words removal etc.
   - The authors have used n-gram, bi-gram, uni-gram methods, TF/IDF, precision and recall. AND IN THE END JUST COMPARED THE RESULTS.

Viewed and worked on different languages like Chinese, English, Spanish, German, Swedish, Romanian, French, Japanese. Publications on Sentiment analysis started from 2004. And today it is one of the fast-growing field in research area of computer science. it has been observed that with the passage of time citations have been increased per year. And it might increase in next few years too

Table 1. Review Dataset description. Total 20285 reviews.

Review sentiments	Training set instances	Test set instances
Neutral ($\mathcal{N}$)	7119	1780
Positive ($\mathcal{P}$)	4880	1220
Negative ($\mathcal{E}$)	4228	1058
Total	16227	4058

Points from Methodology:

- Different query terms are used on DarazPK portal and then the reviews of the users are stored in raw text file.
- Reviews are represented by vector space model (VSM), a.k.a bag of word model. In VSM, each review is represented by the normalized frequency of the words, known as term frequency (TF), and inverse document frequency (IDF) is widely used to find the weights W.

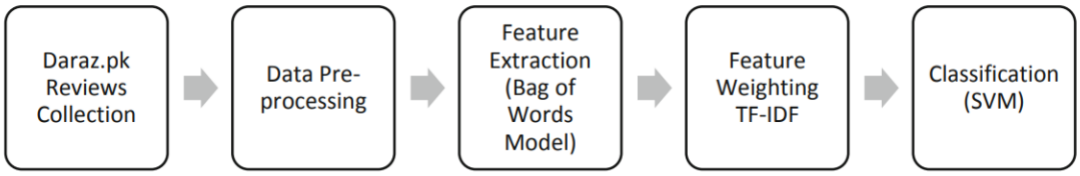


Fig. 1. Basic flow of the framework

- There are total **20285** reviews and There are total **13662** words considered during the experiments.
- To generate the vocabulary, 25 thousand reviews are randomly crawled from Darazpk, frequency of each word is computed and all words present in at least 80% of the reviews are removed (marked as stop words), such as 'a', 'aaaaa', and 'yar'. Once the vocabulary is chosen, then TF-IDF is applied
- **For example, some of the reviews contains only two words such as 'Bahtareen hai yar' which is labeled as P, in English its mean 'it's wonderful buddy', where the word 'yar' (in English buddy) is treated as stop word and removed.**
- Vector Machine (SVM) classifier is used with two kernels:
    1. linear kernel
    2. cubic kernel.
- SVM is inherently a binary classifier. To enable SVM to classify multiple classes, one of the following two approaches are used, namely; (i) one-vs-one (OVO) (ii) one-vs-all (OVA).
- Commonly, OVO approach is better in accuracy than OVA.

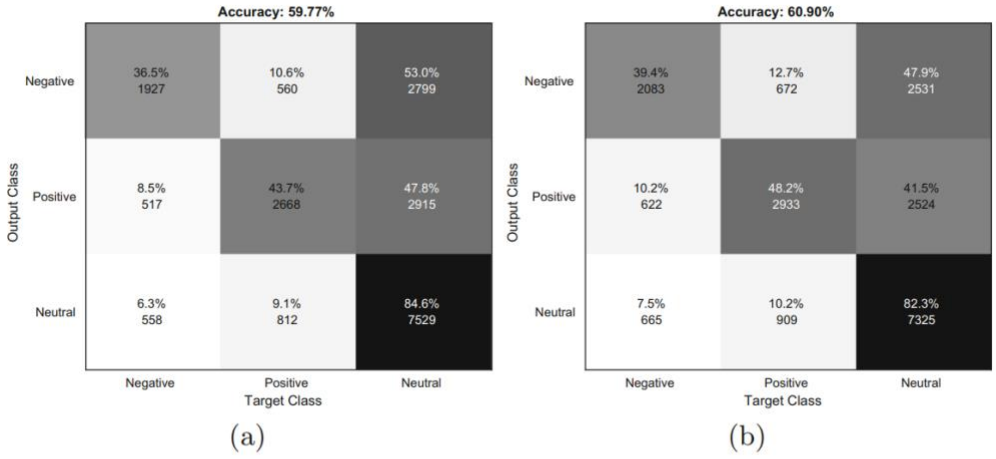
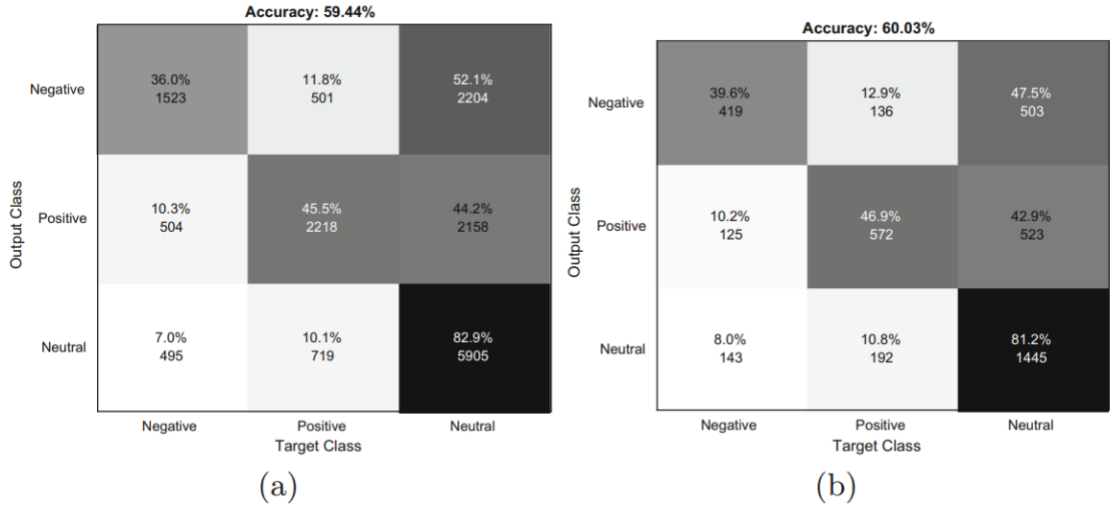


Fig. 2. SVM accuracy on whole dataset. (a) shows the Linear kernel, and (b) shows the cubic kernel.

Experiments and results:

- In literature, cubic kernel outperforms linear kernel in multi-class classification, but in current situation, it is just marginally better.
- In both kernels, the accuracy of sentiment Neutral is very high, whereas the sentiment Negative remains challenging. It is because there are number of ways to show negative feelings for any product, where natural intelligent person can identify but it remains very difficult for artificial intelligent algorithm, unless complex and advance algorithms are not applied such as semantic analysis using natural language processing.
- Below shows the accuracy of cubic kernel, when the dataset is split into training set and test set;



It is interesting to observe that Neutral has highest accuracy and it has highest false positive scores against rest of the sentiments.