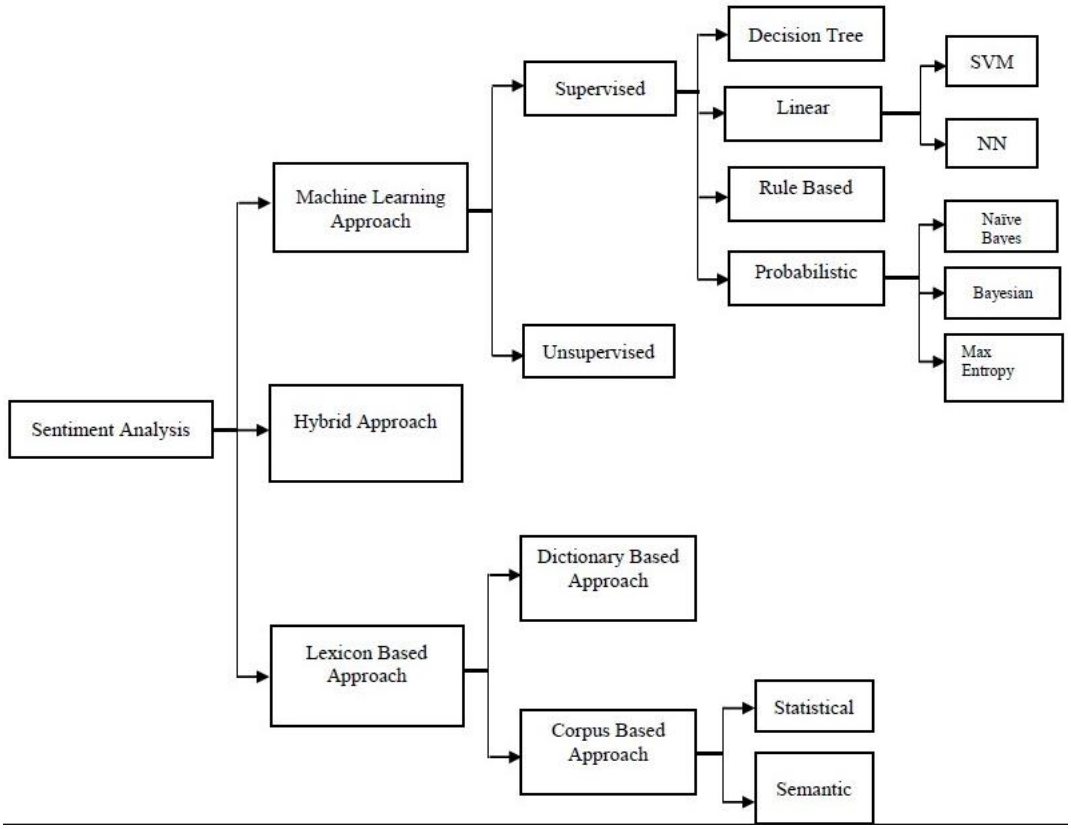
Sentiment Analysis Overview

Approaches



Methodology

1. Data gathering

Very few labelled datasets and corpus is available for Urdu. Data needs to be scraped from online news websites and social media.

2. Data pre-processing

Noise Removal, Named Entity Recognition, Sentence Boundary Detection, Words Tokenization, removing Stop Words, removing punctuation (and symbols, emojis). In some cases, Part of Speech Tagging, spelling correction, lemmatization and stemming can be used.

3. Feature Extraction

1. Ngrams
   Using N-pair of words (monograms, bigrams, trigrams).
2. Syntatic Features
   These are phrase patterns that help detect sentiments.

Refaee and Rieser (2014) used the followingsyntactic features: n-grams of words and POS tags, lemmas, including Bag of Words (BOW), and Bag of lemmas. On the other hand, Al-Sabbagh and Girju (2012) used transitive vs. intransitive verbs. Abbasi et al. (2008) used

3. POS features
   Arabic, there have been several POS tagging approaches including whole word and segmentation-based tagging which means the tagging of the different segments of the word (Mohamed and Ku¨bler 2010). In the latter approach, POS tags contain information about the morphology of the word. POS tagging has been widely used in Arabic text analysis. For instance, El-Halees (2012) and El-Makky et al. (2015) used POS tagging features such as nouns and proper nouns to examine sentiment in Arabic text. Moreover, tools have been made available to find POS tags in Arabic text. One of the tools is MADA toolkit which has been successful by El-Makky et al. (2015) and Habash et al. (2009) in the context of Arabic sentiment analysis.
4. Semantic Features
   These include contextual features which represent the semantic orientation of the surrounding text. Therefore, if an entity has never appeared in the dataset before with the relations of the larger group of entities, the polarity can be detected (e.g. a new iPhone release when the apple product reviews are mostly positive will most likely be positive). Saif et al. (2012) used semantic features to map between entities and their groups. They found that semantic features outperform the unigram and POS tagging features.
5. Lexicon-derived features
   There are many other features that could be added to increase model performance. One example is lexicon derived features such as polarity averages or sums have been commonly used (Zhang et al. 2011; Lu and Tsou 2010).

4. Analysis Application

a) Traditional Machine Learning Algorithms

Following research has been done in Arabic:

Shoukry and Rafea (2012) performed sentiment analysis on 4000 tweets from different domains. They found that 1000 of these tweets were relevant and held opinion without sarcasm. They used two human annotators and found that 500 of the reviews were positive and 500 were negative. They preprocessed the text by removing user-names, pictures, hashtags, URLs, and non- Arabic words. The authors used unigrams and bigrams as features. Pertaining to the machine learning method, SVM and NB were chosen. They performed two experiments: one that included stop words and another that did not. The authors found that removing the stop words led to very small improvement in the performance, indicating that removing stop words add little value to the sentiment in text. SVM performed better than NB by around 4–6% accuracy, with a rate of 72% for unigrams. As for the features, using bigrams did not enhance the results of the unigram model.

Ain et al. (2017) found in their survey comparing different Arabic sentiment analysis studies that deep learning networks are better than SVM and normal neural networks due to the multiple layers that they have. They also stated that deep learning networks have the capability to provide training in both supervised and unsupervised ways.

The naive Bayes classifier reached 96.6% for the Movie dataset when correlated features were used. Character n-grams improved both the SVM and K-NN classifiers accuracies in the Movie dataset which resulted in 89% accuracy. Word n-grams increased the accuracy of K-NN which was 90% for the Movie dataset.

b) Deep Learning Approach

Deep learning models are good as feature engineering is not needed. According to Singhal and Bhattacharyya (2016), deep learning, when given enough data and training time, allows sentiment analysis to analyse data with little restrictions to the specificities of the task or data at hand.

c) Lexicon-based Approach

This approach is mostly used when data is unlabeled.

d) Hybrid Approach

Higher performance than both approaches independently. Lexicon scores are usually used as input to the ML classifier.

Current Urdu Resources

The following is a summary of the research done in Sentiment Analysis in Urdu.

TABLE I. SUMMARY OF EXITING WORK ON URDU SENTIMENT ANALYSIS

Authors	Year	Task	Model/Approch	Polarity	Data scope	Data set/source	Language
Afraz Z. Syed	2010	Lexicon Based SA	Classification	Pos/Neg	Urdu Web Forums	Movies and Products reviews	Urdu
Afraz Z. Syed	2011	Adjectival Phrases as the Sentiment Carriers	Classification	Pos/Neg	Urdu Web Forums	movies and electronic appliances	Urdu
Faiza Hahim	2011	Lexicon Based SA	Classification	Pos/Neg/Neutral	Urdu News Headlines	Product and Movie reviews	Urdu
Smruthi Mukund	2011	identify Opinion Entities	SVM/Kernels Method	N/A	Urdu News Headlines	BBC Urdu News Portal	Urdu
Smruthi Mukund	2012	Analyzing Urdu Social Media for Sentiments	SVM	Pos/Neg	Newswire data	cricket and movies	Urdu
Syed Afraz Z	2014	Identification and Extraction of Appraisal Expressions	Classification	Pos/Neg	Urdu Web Forums	Movies and Products reviews	Urdu
Misbah Daud	2015	Opinion Mining System	Machine Learning	Pos/Neg/Neutral	Roman Urdu	1620 comments	Roman Urdu
S. Abbas Ali	2016	Salience Analysis of NEWS Corpus	Heuristic Approach	Pos/Neg	whole News except the heading	Urdu News Corpus	Urdu
Muhammad Bilal	2016	Sentiment classification	Classification: Models used are Naïve Bay, Decision Tree (DT) and KNN	Pos/Neg	Roman-Urdu and English	The model performance was evaluated on dataset of 150 positive and 150 negative reviews	Roman Urdu
A. Nazir	2017	Opinion Extraction	lexicon-based approach	Pos/Neg	Urdu Web Forums	100,000-tagged words downloaded from http://www.cle.org.pk	Urdu

Centre for Language Engineering (UET)

Center for Language Engineering (CLE) is conducting research and development in linguistic and computational aspects of languages, specifically of Pakistan and developing Asia. Resources on Urdu Normalization, Spell Checking & Translation. They also have POS tagger & annotator software's for Windows. They also have Urdu Wordlists which need to be bought.

Conclusion

Currently, team Abletech has developed code to use labelled Urdu sentences and run various machine learning algorithms on it. Various text-preprocessing techniques can be applied such as removing stop words & making word associations (stemming, lemmatization), however, we require **tagged** (positive/negative) data in order to test the code. These can either be bought or tagged manually for each category.

References

1. https://www.researchgate.net/publication/327985753_Urdu_Sentiment_Analysis
2. https://www.researchgate.net/publication/320738427