Pashto Sentiment Analysis Using Lexical Features

Uzair Kamal
Department of Computer Science, Bahria University
Islamabad, Pakistan
uzairkamal@yahoo.com

Imran Siddiqi
Department of Computer Science, Bahria University
Islamabad, Pakistan
imran.siddiqi@gmail.com

Hammad Afzal
National University of Science and Technology
Islamabad, Pakistan
hammad.afzal@mcs.edu.pk

Arif Ur Rahman
Department of Computer Science, Bahria University
Islamabad, Pakistan
badwanpk@gmail.com

ABSTRACT

Individuals use various platforms to express their opinion regarding products, services, political situations and other events. Knowing the opinion of people is very important for the concerned individuals and organizations in order to devise future strategies according to the wishes of people. The present research study focuses on extraction of opinion from digital-born Pashto text. The study involved the creation of multiple state-of-the-art classifiers by adapting methodology of message level task using sentiment analysis of Tweets'. In addition to this, word-sentiment lexicons with tokenization of sentences and translation of existing English lexicons were generated. The findings show that lexical features based Pashto sentiment analysis extracts sentiments with a high accuracy.

CCS CONCEPTS

•Computing methodologies →Lexical semantics; Language resources;

KEYWORDS

sentiment analysis, lexical features, Pashto text

ACM Reference format:
Uzair Kamal, Imran Siddiqi, Hammad Afzal, and Arif Ur Rahman. 2016. Pashto Sentiment Analysis Using Lexical Features. In *Proceedings of MedPRAI-2016, Tebessa, Algeria, November 22-23, 2016,* 4 pages.
DOI: http://dx.doi.org/10.1145/3038884.3038904

1 INTRODUCTION

In the recent years, individuals use diverse platforms to express their opinions and sentiments in the form of blogs, newspaper columns and social networking websites. The opinions of people are of very important for governments and organizations to get the feedback about their policies, products and services in order to eradicate the issues and devise improvements. Manually extracting opinions by scanning blocks of text becomes a expensive, tiresome and tedious task because of the rapid growth of data. Consequently, a number of computerized systems to automatically classify opinions from given text have been researched and developed [1] [5]. Research on automatically extracting opinions from born-digital text is pretty mature on many languages spoken around the globe. These mostly include languages which use Latin alphabet like English, Spanish and French. However, the state-of-the-art is still not very pleasing when it comes to complex morphological languages like Arabic, Urdu and Pashto [16] [15]. In addition, the traditional methodologies based on lexicon, parts of speech (POS) tagging and Named-Entity (NE) recognition fail when applied to such languages.

The present research targets Pashto language which is the second largest regional language spoken in Pakistan and has the status of official language in Afghanistan. Multiple sentiment models with inputs of lexical features created from the lexicon presented in [10] are developed. Pashto lexicons were generated manually with predefined rules [14] which can classify sentences with an accuracy of 73.2%. Consider, for instance, the following two sentences in Pashto.

(1) "زه تاسو لپاره درناوي لرم، تل ده غوره کار کولو"
(I greatly admired your excellent work.)

(2) "نوکیا اواز کیفیت ډهیر بد وي"
(The sound quality of Nokia is very bad.)

The first sentence portrays positive sentiment while the second shows a negative sentiment. Words like "درناوي" and "غوره" define positive sentiment of the sentence, while the

MedPRAI-2016, Tebessa, Algeria

DOI: http://dx.doi.org/10.1145/3038884.3038904

while the word "بد" portrays the negative opinion in the sentence. This research targets such words and generates lexical features from such annotated words which help in defining sentiment for text.

There are strict inflectional or derivations rules on morphemes in English text. For instance, suffixes like 's and 'es are used to make plurals with exception of a few words. Pashto, on other hand, has a very complex morphology. Pashto plural morphemes have several all-morphs e.g. plural of "هلك" (boy) is "هلکان" (boys), for the word "ګل" (flower) plural is "ګلونه" (flowers), "پیغلې" (girls) for "پیغله" (girl) and there are many more such examples. This makes analysis of Pashto text more challenging as compared to other languages.

User comments and posts from various social media and news websites including Facebook [1], BBC [2], Twitter [3], VOA [4] were collected for the study. Each of the comments was manually labeled as either positive or negative to generate the labels. A lexicon was extracted from the downloaded user's comments and a second one was created from translation of existing English lexicon to Pashto. Classifiers including Support Vector Machine (SVM), Logic-based Machine Translation (LMT), NaiveBayes and J48 (C4.5) are evaluated. The details of the proposed technique are presented in the following section.

2 PROPOSED METHODOLOGY

The section details the steps followed to develop the system for extracting sentiments from born-digital Pashto text.

2.1 Corpus

The most challenging part of the research was to collect an opinion enriched corpus of Pashto. Most of the Pashto speakers at social media express themselves in Roman Pashto (transliterated) which is not standardized. Automatic retrieval along with manual collection was used to retrieve users comments from different sources like Facebook, BBC, Twitter and VOA as mentioned earlier. A total of about 600 sentences were collected to build the corpus out of which 100 were reserved for testing.

2.2 Pre-processing

This phase prepares sentences for lexicon creation and sentiment analysis. Sentences contain different URLs, differently written same words and HTML tags extracted from News websites and other similar sources. In Pashto, Urdu and Arabic, unlike Latin languages, words are not usually separated by empty spaces. These morphologically complex languages therefore face problems like space insertion and space omission errors [4]. Being the premier research on Pashto text, pre-processing part in our case is limited to filtering of URLs and non-Pashto sentences. Corpus normalization was carried out manually by saving different forms of the words and assigning them same polarity and score.

2.3 Tokenization

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens (string with meaning). It is carried out by the Stanford CoreNLP [5] tokenizer, which splits sentences into tokens.

2.4 Lexicon Creation

Lexicon building comprises of two phases.

- New Lexicons: All tokens from tokenization phase are iterated manually and assigned a polarity and a score. In the present research work, the score assigned to all tokens is one while the polarity is positive or negative. All the words which neither have a polarity nor a score are ignored while building lexicons.
- Existing Lexicons: Lexicons from English language used by Bing and Liu [3] and by Nielsen [11] are translated from English to Pashto and assigned polarity. The translation was using a semi-automated approach i.e. google translate was used but in some cases the the translation was manually improved.

Consider, for instance, the following two Pashto phrases.

(1) "زه تاسو لپاره درناوي لرم، تل ده غوره کار کولو"
(I greatly admired your excellent work.)

(2) "نوکیا اواز کیفیت ډهیر بد وي"
(The sound quality of Nokia is very bad.)

Adjectives like "غوره" and "بد" are lexicon candidates, they are assigned the respective polarity and the same score as shown in Table 1.

[1] https://www.facebook.com/bbcpashto1981/?fref=ts
[2] http://www.bbc.com/pashto
[3] https://twitter.com/bbcnewspashto
[4] http://www.voadeewaradio.com/
[5] http://stanfordnlp.github.io/CoreNLP/

Table 1: Lexicons

Lexicon	Polarity	Score
غوره	positive	1
درناوي	positive	1
راحت	positive	1
وزلي	negetive	1
بد	negetive	1
دوښمني	negetive	1

This step results in a list of about 2500 lexicons which are used in generation of Lexical features.

2.5 Features and Classifiers

Each sentence is tokenized and the sentences are represented by feature vectors from unigram tokens with the lexical feature generated for each token [10]. For each token w and emotion or polarity p, we used the sentiment/emotion score $score(w, p)$ to determine the following.

- Total count of tokens in the tweet with: $score(w, p) > 0$;
- Total score = $\sum_{w \in tweet} score(w, p)$
- Maximal score = $maxw(w, p)$;
- The score of the last token in the tweet with: $score(w, p) > 0$;

Four lexical features using the above formulas are constructed for each polarity. The positive features are represented by PositiveF1, PositiveF2, PositiveF3 & PositiveF4 and the negative features are represented by NegativeF1, NegativeF2, NegativeF3 and NegativeF4. For each sentence these features are added to a feature vector in training file as shown in following example.

```
@attribute PositiveF1 numeric
@attribute PositiveF2 numeric
@attribute PositiveF3 numeric
@attribute PositiveF4 numeric
@attribute NegativeF1 numeric
@attribute NegativeF2 numeric
@attribute NegativeF3 numeric
@attribute NegativeF4 numeric
@attribute class {Positive,Negative}
```

For example, a given sentence without Lexical features inclusion can be represented as follow:

```
@data
{4 0,6 0,7 0,39 0,59 0,140 0,
322 0,865 0,1501 0,1502 0,
1503 0,1554 Negative}
```

While sentence with inclusion of Lexical features:

```
@data
{4 0,6 0,7 0,39 0,59 0,140 0,
322 0,865 0,1501 0,1502 0,
1503 0,1558 1,1559 1,1560 1,
1561 1,1563 Negative}
```

A Support Vector Machine(SVM) model using LibSVM is trained for classification [8] [2]. A linear kernel with C=0.005 is applied. Models were also trained with Naive Baye's and Logic-based Machine Translation (LMT) [6], J48 [13] [9].

3 EXPERIMENTS AND RESULTS

The experiments are carried out on the developed corpus of annotated sentences with equal distribution of positive and negative examples. Moreover, we also generated 2500 lexicons and assigned a score of 1 to all. Training files were created with Weka [7] 'arff' file format using Java. Lexical features [10] for positive and negative examples were created and added to 'arff' file during training.

As mentioned earlier, we employed supervised learning with uni-gram approach [12] and generated four different binary classification models with SVM, LMT, Naive Bayes and the J48 classifier.

We present the system accuracy in terms of F-score for each of the models with and without lexical features for different sizes of the training data set. The experiments are carried out using k-fold cross validation, the number of training sentences is varied for different experiments while the test set is fixed to 100 sentences. Table 2 summarizes these results for the SVM model while Table 3 presents the results for the Naive Baye's classifier. Likewise, Table 4 and Table 5 summarize the system performance for J48 and LMT classifiers respectively.

Table 2: Performance using SVM

Sentences	Features	F-Score
200	WithoutLexicalFeatures	0.60
200	WithLexicalFeatures	0.655
300	WithoutLexicalFeatures	0.647
300	WithLexicalFeatures	0.694
500	WithoutLexicalFeatures	0.686
500	WithLexicalFeatures	0.732

Table 3: Performance using Naive Bayes

Sentences	Features	F-Score
200	WithoutLexicalFeatures	0.58
200	WithLexicalFeatures	0.67
300	WithoutLexicalFeatures	0.62
300	WithLexicalFeatures	0.67
500	WithoutLexicalFeatures	0.69
500	WithLexicalFeatures	0.71

Table 4: Performance using J48

Sentences	Features	F-Score
200	WithoutLexicalFeatures	0.59
200	WithLexicalFeatures	0.61
300	WithoutLexicalFeatures	0.60
300	WithLexicalFeatures	0.63
500	WithoutLexicalFeatures	0.65
500	WithLexicalFeatures	0.69

Table 5: Performance using LMT

Sentences	Features	F-Score
200	WithoutLexicalFeatures	0.63
200	WithLexicalFeatures	0.62
300	WithoutLexicalFeatures	0.63
300	WithLexicalFeatures	0.69
500	WithoutLexicalFeatures	0.67
500	WithLexicalFeatures	0.72

Comparing the performance across different classifiers, it is interesting to observe that the performance is more or less consistent across different classifiers. The performance naturally increases with the increase in the size of training corpus. A highest F-score of 0.732 is observed using the SVM classifier with lexical features when using a training corpus of 500 sentences.

4 CONCLUSION AND PERSPECTIVES

This paper presented the first attempt on sentiment analysis on text in Pashto language. The major contributions of this study include building a Pashto corpus, lexicon list and a generalized classification framework based on lexical features to predict sentiments from a given text. The present study is based on uni-gram approach and can be enhanced further by introducing the n-gram approach. We also intend to enhance the corpus and make it publicly available.

REFERENCES

[1] Mohammad Ehsan Basiri, Ahmad Reza Naghsh-Nilchi, and Nasser Ghassem-Aghaee. 2014. A Framework for Sentiment Analysis in Persian. *Open Transactions on Information Processing* 1, 3 (November 2014), 1–14.

[2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A Holistic Lexicon-based Approach to Opinion Mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 231 – 240. DOI:http://dx.doi.org/10.1145/1341531.1341561

[4] Nadir Durrani and Sarmad Hussain. 2010. Urdu Word Segmentation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 528–536. http://dl.acm.org/citation.cfm?id=1857999.1858076

[5] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. 2014. Sentiment Analysis in Arabic tweets. In *2014 5th International Conference on Information and Communication Systems (ICICS)*. 1–6. DOI:http://dx.doi.org/10.1109/IACS.2014.6841964

[6] Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian Network Classifiers. *Mach. Learn.* 29, 2-3 (Nov. 1997), 131–163. DOI:http://dx.doi.org/10.1023/A:1007465528199

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. DOI:http://dx.doi.org/10.1145/1656274.1656278

[8] Marti A. Hearst. 1998. Support Vector Machines. *IEEE Intelligent Systems* 13, 4 (July 1998), 18–28. DOI:http://dx.doi.org/10.1109/5254.708428

[9] Michael C. McCord. 1989. Design of LMT: A Prolog-based Machine Translation System. *Comput. Linguist.* 15, 1 (March 1989), 33–52. http://dl.acm.org/citation.cfm?id=68960.68963

[10] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *CoRR* 2 (June 2013), 321 327. http://arxiv.org/abs/1308.6242

[11] F. Å. Nielsen. 2011. AFINN. Informatics and Mathematical Modelling, Technical University of Denmark. (March 2011). http://www2.imm.dtu.dk/pubdb/p.php?6010

[12] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. DOI:http://dx.doi.org/10.1561/1500000011

[13] Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

[14] I. Rabbi, M. A. Khan, and R. Ali. 2008. Developing a tagset for Pashto part of speech tagging. In *2008 Second International Conference on Electrical Engineering*. IEEE Xplore, Lahore, Pakistan, 1–6. DOI:http://dx.doi.org/10.1109/ICEE.2008.4553909

[15] Kashif Riaz. 2008. Concept Search in Urdu. In *Proceedings of the 2Nd PhD Workshop on Information and Knowledge Management (PIKM '08)*. ACM, New York, NY, USA, 33–40. DOI:http://dx.doi.org/10.1145/1458550.1458557

[16] Afraz Z. Syed, Muhammad Aslam, and Ana Maria Martinez-Enriquez. 2010. Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. In *Proceedings of the 9th Mexican International Conference on Advances in Artificial Intelligence: Part I (MICAI'10)*. Springer-Verlag, Berlin, Heidelberg, 32–43. http://dl.acm.org/citation.cfm?id=1927149.1927155