Developing Roman Urdu Lexicon and Corpus for Sentiment Analysis

Muhammad Naeem Ahmed Khan, Asad Mehmood

SZABIST, Islamabad, Pakistan

Abstract
There have been a large amount of quality work done in the area ofsentiment analysis for opinion mining for the social media content available in English language, but no such work is reported in the contemporary literature for the Roman Urducontents. Most of the Tweets which have been posted on social media from Pakistan are mixture of English and Roman Urdu. A number of Tweets are made in Roman Urdu and when it comes to sentiment analysis those Tweets are simply ignored; hence putting a question mark on the correctness of the results produced from the selected Tweets. In this paper, we have made an attempt to build a Semantic Lexicon and Roman Urdu Corpus duly tagged for positive and negative sentiments. This can also be used in future for automated Lexicon extension and mining sentiments from the Tweets.

Keywords:
Sentiment Analysis, Opinion Mining,Semantic Web, Twitter Analysis, Roman Urdu Corpus, Hash Tags, Big Data Analytics.

1. Introduction

The fad of microblogging has now become popular among the Internet users. Microblogging is one of the broadcast medium which is an underlying form of blogs. Microblog has smaller actual and aggregate size and it is very different in relation to traditional blogs. These tiny microblogs are like short sentences, images or video links, and are also known as microposts (Kaplan & Haenleinm, 2011; Lohmann et al., 2012) .Millions of people across the world use this medium to express their views on different daily life events and prospects of. Since, this trend has emerged recently; a lot of approaches related to sentiment analysis have been devised and explored.

Internet users use blog posts and forums to express their views about different products, services or issues. Sentiment analysis performed on the contents available on these platforms may provide useful information to the security analysts, but the key challenge is how to perform such analysis on such a big data.Twitter is one of the evolving social media which,on average, hosts around 200 million tweetsperday. Tweets can correspond to innumerable products, services, social issues, news, incidents and reviews etc. Further, people also comment and share views about tweets pertaining to varioussocial and political topics/issues. Twitter shares these tweets in a unique way, which cannot out-rightly be used to judge the essence of the tweets and topics being discussed on social media.

Social media websites contain a large amount of data on almost every type of discipline, subject and issue. Suchtype of data or information is mostly found in the form of comments and feedback, reviews and criticism etc. Such information also proves to be of significant importance for companies and groups which market their products and promote them on the Internet. However, it is cumbersome to analyze all of this data. Hence, it becomes imperative to devise an appropriate methodology or technique to automatically process and analyze the opinions or sentiments present within this data. This paves a way for inducting natural language processing, computational linguistics and text mining techniques to analyze social media content.

Twitter is one of the evolving social media that acts as an instant communication source. The real-time information supplied by millions ofusers is readily available on Twitter all the time. Different organizations rely on such information to analyze and understand the customers' views about their product or service. For example, TV channels use Twitter tweets as the source of feedback about their programs and talk shows to measure their viewership ranking. Although a single tweet consists of maximum 140 characters, but this information becomes large due to numerous tweets being made by thousands of users on a single issue within a shorter span of time. Eventually, analyzing such a large amount of data, which is totally in unstructured text form, becomes a colossal task. Hence, the need for an automated classifier becomes necessary to reduce the time for analyzing the data and improve efficiency of the analysis process.

Social media such as Facebook, Twitter etc. is one of the fast evolving phenomena for sentiment analysis using various applications. In the present day technology driven culture, we can get opinions from different polls and advertisements placed on blogs and social media sources. In general, humans have an instinct to share information or give feedback about the products they purchase in their daily lives. And this trait of sharing has now moved onto social media sites like Twitter and other microblogging sites and sources. By this means, these sources are becoming

Manuscript received July 5, 2026
Manuscript revised July 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.7.8

very useful in identifying and analyzing diverse opinions on different topic and areas. Twitter is an important source for getting opinion from microblogging data available in different languages. Such kind of data can be obtained by using the Twitter's "Tweet Entities".

Only those tweets whose visibility is set as public can be fetched from Twitter. A tweet cannot be more than 140 characters and this characteristic makes it somewhat easy to analyse numerous tweets easily. Prior to analyzing data for sentiment analysis and opinion mining, it is necessary to preprocess it using various techniques as mentioned in (Gokulakrishnan et al., 2012). The term "Emoticon" refers to the emotion icon used in tweet like smiley and sad gestures icons can be expressed by using ":)" and ":(" symbols respectively. Therefore, it is necessary to replace such symbols by suitablekeywords to identify them as emoticon. In *Uppercase Identification*, the uppercase words are replaced by some identifiable keyword. *Lower Casing* is done to lowercase the whole tweet to maintain consistency in the dataset. Using *URL Extraction*, all URLs present in a tweet are replaced with a tag to identify it as a URL in order to reduce size of the tweet. This uppercase and lowercase conversion can also be done through "toupper()" and "tolower()" functions. To further reduce size of the tweets, usernames and hash tags are reduced to some constant symbol and this conversion is known as *Detection of Pointers*. In addition, punctuations marks are also removed to avoid inconsistency in training dataset. In *Removal of Query Term*approach, the query term is used to fetch the required Tweets. Hence, the query termsare useless to identify the sentiments in the post and they are replaced with the tag. Similarly, *Compression of Words* is done for the elongated words by compressing the word having three repeated letter e.g., the word "cooooool" is replaced with "cool" so that it can be differentiated between the regular usage and emphasized usage.

After preprocessing the dataset, the sentiment analysis of the Twitter data can be carried out using different classifiers. The *Naïve Bayes* classifier is based on the Bayes' theorem and probabilistic model (Murphy, 2006). *RandomForest* constructs different genre of classification trees. A tree tells about the class for the particular input vector and the class with high score is selected (Breiman, 2001). The accuracy and high score depends upon the how much independent it is from the other trees. *Support Vector Machines (SVM),*termed as non-probabilistic binary linear classifier,is a popular classifier that takes input vector of tweets and determinesits possible class. *Sequential Mining Optimization (SMO)* algorithm is used for the optimization problems when training support vector machines. SMO solves the problem step by step and splitsthe problem into smaller parts to solve and analyse it analytically. *J48* is an open source java implementation of C4.5 algorithm used to generate decision trees from the dataset.

There is a large amount of data available over the Internet on which users provide diverse opinions about different contentlike blogs, social networks, web forums etc. The major chunk of Internet contains the subjective opinions of certain users which influence the opinions of others on the Internet. We can rightly say that in today's social media environment each individual has become a propagandist. Due to the presence of such content over the Internet, we need to develop methods which automatically analyze subjective aspect of the content.

Therefore, software industry is focusing more on building applications which are based onopinions as there is a large amount of consumer generated information available in the form of blogs, message boards, forum and news articles which providesenormousopportunity to understand the opinions about the products.In case the companies do not pay attention to these types of contents, then they run a chance to lose their brand image or name.Sentiment analysis can also prove to be very helpful for analyzing the stock market trend. The investors require understanding the stock market news instantlyto capture stock market trend so that they can invest or disinvest shares in stock market in a timely manner.

The key question arises, why we need sentiment analysis. Sentiment analysis basically tries to look into different aspects of natural language which help people find valuable information from large amount of unstructured data (Bloom et al., 2007). It is an emerging concept in which different human emotions are detected from textual content, thus enabling us to extract opinions and sentimental feelings of the people on a subject or issue. It is really hard for the computing machines to automatically extract the exact meanings and tonesfrom the document content as people express so many things in several different ways and styles. These sentiment analyses prove very useful when we analyze search engine results, different blogs, social networks, web forums, different review of people on books, movies, sport and products etc (Hasan & Adjeroh, 2011). These reviews help people analyze opinions and features about the different products and services.In the recent times, the social networks have become very popular and famous amongst the people all over the world. People talk about products, services, social issues, entertainment, politics and many other topics. If there is a system which can perform the sentiment analysis on these areas, then it would help find out the sentiments and opinions expressed in textual information.

2. Literature Review

In order to evaluate large volume of Twitter data, the following three techniques have been introduced by Hao *et al.* (2011) which are mainly based on visual sentiment analysis:

(i) *Topic-based analysis*employs natural language processing techniques to extract certain attributes to determine nature of the topic of discussion. The technique detects different opinion-related attributes to measure the degree of sentiment;

(ii) *Stream analysis* is performed on large number of tweets. It extracts popular and interesting tweets based on negative/positive attitude and the influencing characteristics that it possesses within the larger tweetsdensity;

(iii) *Pixel cell-based sentiment calendars and high density geo-maps* are used to visualize and depict large number of tweets in a single view. This technique when applied to a large volume of tweetshelp investigate how tweets are distributed and to verify the opinions that influence audience the most.

Twitter being the sources of microblogging platform is commonly used for sentiment analysis. Pak & Paroubek (2010) used English language tweets corpus for sentiment analysis and opinion mining. The corpus containing emotions like smiley "*:)*" or sad "*:(*" was readily evaluated as positive or negative sentiments respectively. The corpus was analyzed to check how data was distributed into subjective corpus (containing positive or negative set) and objective corpus (containing neutral set). The analytics showed that objective sets contained more common and proper nouns which in turn had often used personal pronouns. Similarly, the objective sets bloggers addressed themselves as a third person while the subjective sets bloggers described themselves as the first person or second person. The authors calculated n-grams for extracting binary features and keyword frequency was used to obtain rest of the general information.

Li & Liu (2010) applied clustering-based sentiment analysis approach on dataset which was based on the review of a movie. The dataset was divided into two primary clusters corresponding to the positive and negative aspects of the users' comments and remarks about that movie. For experimental purpose, the technique of TF-IDF (Term Frequency – Inverse Document Frequency) was applied. The stability of results was increased when the voting mechanism was used, and finally a symbolic technique was applied to enhance the clustering results. In symbolic technique each term (or word) is assigned some value based on the negative and positive connotation and intensity of that term. Then, an aggregation functions is used to obtain cumulative scores to draw conclusions about the overall negativity or positivity of the comments.

Lima & de Castro (2012) propose an automatic classifier for Twitter messages that comprises three modules: Support Counting, Database Selection and Classification modules. *Support Counting* module counts percentage of the tweets that contain at least one word or emotion in the tweet. *Database Selection* module divides data into two sets: training set and testing set. *Classification* module classifies data using Naïve Bayes algorithm. The authors used three approaches for automatic sentiments classification. Firstly, e*motions-based approach* is used to evaluate emotional sentiments which have been expressed in tweets. Then, w*ord-based approach* sifts those words that describe the sentiments for automatic classification. Finally a hybrid approach is employed that uses both the emotions and word-based approach to classify the data. In this way, the classifier checks whether the testing dataset satisfies the minimum threshold set by the system.

Another system based on the estimation of social media sentiment is proposed by Colbaugh & Glass (2011). The author proposed the system in which they formulate the learning-based text classifier. The classifier just models the problem data in such a way that the data is shown as the bipartite of documents and words Colbaugh & Glass (2011). The authors made some suppositions and assumptions, in which they assumed that the small lexicon of words will be used by the sentiment classifier for learning which is called sentiments orientation. Furthermore, it was supposed that the non-labeled documents will be used for training. The proposed technique does not need to collect and label the document and each time the classifier is retrained whenevera new document or data becomes available for analysis.

A novel method for sentiment analysis, called proximity-based sentiment analysis, is proposed by Hasan & Adjeroh (2011) in which new features based on word proximities are considered. The authors used three proximity-based features which are called proximity distribution, mutual information between proximity types and proximity patterns. Initially, the dataset was cleansed and contained equal number of negative and positive reviews from different movies. A single dataset was divided into number of segments in which each segment contained over 100 words. The distance between positive and negative pair of word was calculated. Once the word proximities of different words had been calculated, the author used three proximity-based features in order to extract the pair-wise information. The three proximity-based features which were used include:*Proximity Distributions*(different numbers of bin were considered which returned the distribution of pair-wise distances from the proximity models), *Mutual Information between Proximity Types* (the relationship between the proximity types was used to determine the polarity of the document; and the theoretic quantities of entropy for each sequence was used to get mutual information between pairs of proximity types), and *Proximity Patterns* (which described polarity of words used in document). Finally, the authors applied three different classification techniques on the extracted proximity

features.In another research on sentiment analysis, Koncz & Paralic (2011) proposed a different approach for feature selection. The authorscoined a formula for the calculation of attribute weight. The approach was to create a method which has higher accuracy in comparison with the other methods. In the proposed method, the feature selection approach usesfrequencies of document terms for particular categories. The values are normalized in terms of total number of documents in that category. The formula used by the author assigns a higher value to those attributes which have lower frequency and those valuesare characterized for the given category.

Mizumoto *et al.* (2012) proposed the system by creating polarity dictionary to determine the sentimental polarities of stock market. While constructing the dictionary, the authors used semi-supervised learning approach from which small polarity dictionary is made. Using the co-occurrence frequency with words in polarity dictionary, only those new words are added to the dictionary for which polarities are unknown. The polarity of article was determined according to the frequency of words in the polarity diction which was created earlier using the semi-supervised learning method. In this way, all the articles are determined as positive, negative or neutral.

Karamibekr & Ghorbani (2012) classify sentiments using verbs from the sentenceby defining a model that keeps relationship between sentiments in the levels of document, sentence and opinion. The approach uses the verb as a core opinion term in the social domains. Since verb is considered as the main opinion term, the authors construct the opinion verb dictionary which had 440 opinion verbs. The authors further use the lexical knowledge to extract the sentiment terms present in the sentence. A subject in the sentence describes the action of the doer or what the predicate does; whereas the object in a sentence is what or whom the verb is acting.With the help of bootstrapping process, a list of opinion verbs is collected using English dictionary and its synonyms in the WordNet. Then, Stanford POS tagger (Toutanova et al., 2003) is used to recognize verbs and extract element of the sentence.

While creating effective sentiment based taxonomy, Cai *et al.* (2008) proposed a statistical based approach for sentiment analysis that does not assume any particular subject or object in the sentence. The sentiments expressed by the words are measured on the scale of positive or negative. Then, a list of positive and negative words is created. In order to score the relative sentiment for the posts which have the positive and negative words, the authors characterize the degree of positivity/negativity that each word conveys. Cho & Kang (2012) introduced an approach for text sentiment classification based on the contextual information obtained from different paragraphs. The authors use sentiment based domain dictionaries to carry out the sentence sentiment classification which covered the formal and informal vocabularies. The information about the flow of sentiments and keywords in the paragraph is taken out from the contextual information. To get the contextual information of the whole paragraph, the author carried out the following process. Firstly, reliable keywords are taken out from domain training data. Then, for each keyword, the importance of weight is based upon the frequency rate in training dataset. Finally, the frequency value of each word is normalized with the weighting factor.

3. Conceptual Model

We searched Roman Urdu Tweets generically as our research was not specific to searching tweetsof a specific user. We used Twitter's SearchAPI to collect the required tweets. During this study, we observed that Twitter cannot fetch Tweets which are older than one week. We obtained data from Twitter for a variety of different words and hashtags based on types of the political and social events being actively discussed on social media those days. We particularly obtained tweets related to cricketmatches andpolitical issues. The detaileddescription of how we gathered the required data is given below.

We searched tweets based on major cities of Pakistansuch as Quetta, Peshawar, Islamabad, Karachi and Lahore. We got the latitude, longitude of the cities from Google Maps API. By selecting a suitable center position foreach of the city, we set the radius range of 25 km from the assumed centre. For instance, the latitude and longitude chosen by us for the assumed centre position of Islamabad were 33.718151^{o}, 73.060547^{o} respectively; and from this centre position, we covered the area of 25 Km radius to gather the tweets originated from Islamabad. In the next step, we preprocess the tweets and URLS, hashtags and usernames were replaced with white space. Subsequently, the multiple spaces were replaced with a single space and the line feed and carriage returns were removed.In the preceding steps, the data was sifted for English Language Tweets and those word which were not found in English dictionary were considered as Roman Urdu. The tweets were analyzed manually and were marked as negative and positive Tweets and the ones which were left over were considered as the neutral (or balanced) sentiment. The specific words due to which the tweetswere marked as positive or negative were stored into the dictionary, with weight of +1 and -1 for positive and negative words respectively.

Our proposed model is provided in Figure 1which consists of two phases; one to create Roman Urdu dictionary and other to analyze the sentiment of English Tweets and Roman Tweet Sentiment to generate the overall combined sentiment.

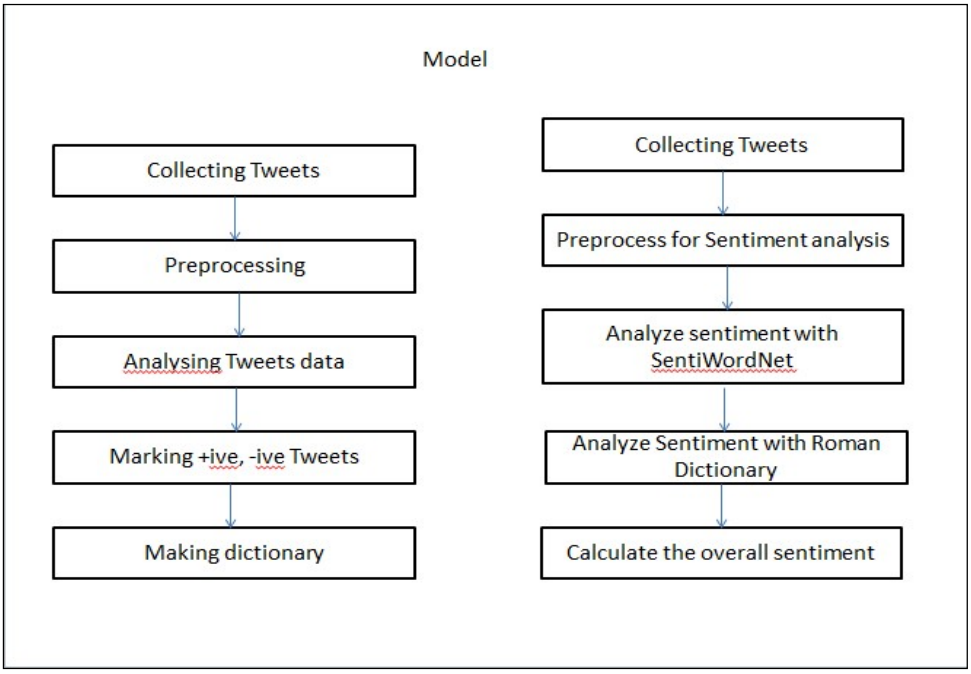

Figure 1: Proposed Model for Sentiment Analysis.

In the initial phase, we started creating the Roman Urdu dictionary. For this purpose, we collectedtweets and then preprocessed them. In the preprocessing phase, the special characters were removed and any other non-English words were also removed from the tweets. The tweet data which contained Roman Urdu tweets were analyzed manually by us. The Roman Urdu tweets were assembled in a manner such that the positive or negative words could be assigned corresponding positive or negative score. The Roman Urdu tweets were not stemmed and major words were marked as positive and negative words.However, it is worth mentioning that the words are marked positive and negative by different people as per their own understanding. To account for this issue and bias amongst different persons' opinions, only one person was detailed to mark the tweet as positive or negative.

In the second phase, we collect tweets from a topic and preprocess them for sifting the English words/sentences.In the preprocessing stage, the special characters were removed and any other non-English words were also removed from the tweets. Furthermore, the stop words (like her,him,above,near,all) are removed and the process of stemming is also carried out after removing stop words. In the stemming stage, the action verbs which describe and action of the subject are turned into 1st form of the verb.Sentiment analysis for English tweets is done with the help of SentiWordNet Dictionary. The scores for each word in the tweet were added to get an overall score.In the next step, the sentiment analysis for Roman Urdu words was carried out with the help of Dictionary which we made in initial phase as shown in Table 1. Also, the scores for each word in a tweet were added to get an overall score for Roman Urdu tweets. In the last step, the sentiment score for English Tweets and Roman Urdu are added in order to generate an overall score of the tweet.

Table 1: An Excerpt Sample of our Roman Urdu Dictionary

Sr #	Roman Urdu Word	Meanings (in English)	Positive/ Negative Connotation	Score
1	Nahi	No	-1	-1
2	Acha	Good	+1	+1
3	Bura	Bad	-1	-1
4	Behtreen	Best	+1	+1

Data	Positive	Negative	Neutral
SentiWordNet	28.8%	15.6%	55.4%

Now, we add both the sentiments to see how the Roman Urdu Dictionary has enhanced the Twitter sentiment.

Data	Positive	Negative	Neutral
Overall	29.4%	16.7%	53.7%

4. Evaluation and Results

The results and calculations are done in such a way that we created our Roman Urdu Dictionary. The tweets were gatheredfor a cricket match between Pakistan and Sri Lankaplayed in 2015. The Tweets were preprocessed and stemmed for sentiment analysis using SentiWordNet English dictionary. The sentiment for each Tweet was calculated and saved into the database. In the proceeding steps, the sentiment was calculated using the Roman Urdu Dictionary for those Tweets.

References

[1] Bloom, K., Garg, N., & Argamon, S. (2007). Extracting Appraisal Expressions. In *HLT-NAACL* (Vol. 2007, pp. 308-315).

[2] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[3] Cai, K., Spangler, S., Chen, Y., & Zhang, L. (2008, December). Leveraging sentiment analysis for topic detection. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*(Vol. 1, pp. 265-271). IEEE.

[4] Cho, S. H., & Kang, H. B. (2012, October). Statistical text analysis and sentiment classification in social media. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on* (pp. 1112-1117). IEEE.

[5] Colbaugh, R., & Glass, K. (2011, September). Agile sentiment analysis of social media content for security informatics applications. In *Intelligence and Security Informatics Conference (EISIC), 2011 European* (pp. 327-331). IEEE.

[6] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on* (pp. 182-188). IEEE.

[7] Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D., Haug, L. E., & Hsu, M. C. (2011, October). Visual sentiment analysis on twitter data streams. In*Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*(pp. 277-278). IEEE.

[8] Hasan, S. S., & Adjeroh, D. (2011, August). Proximity-based sentiment analysis. In *Applications of Digital Information and Web Technologies (ICADIWT), 2011 Fourth International Conference on the* (pp. 106-111). IEEE.

[9] Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, *54*(2), 105-113.

[10] Karamibekr, M., & Ghorbani, A. (2012, December). Verb oriented sentiment classification. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on* (Vol. 1, pp. 327-331). IEEE.

[11] Koncz, P., & Paralic, J. (2011, June). An approach to feature selection for sentiment analysis. In *Intelligent Engineering Systems (INES), 2011 15th IEEE International Conference on* (pp. 357-362). IEEE.

[12] Li, G., & Liu, F. (2010, November). A clustering-based approach on sentiment analysis. In *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on* (pp. 331-337). IEEE.

[13] Lima, A. C., & de Castro, L. N. (2012, November). Automatic sentiment analysis of Twitter messages. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on* (pp. 52-57). IEEE.

[14] Lohmann, S., Burch, M., Schmauder, H., & Weiskopf, D. (2012, May). Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 753-756). ACM.

[15] Mizumoto, K., Yanagimoto, H., & Yoshioka, M. (2012, May). Sentiment analysis of stock market news with semi-supervised learning. In *Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on*(pp. 325-328). IEEE.

[16] Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*.

[17] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 1320-1326).

[18] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.