TWITTER SENTIMENT ANALYSIS FOR ALBANIAN LANGUAGE

Alba HAVERIKU
Faculty of Information Technology, Polytechnic University of Tirana, Albania
Neki FRASHËRI
Albanian Academy of Sciences, Tirana, Albania

ABSTRACT

The large amount of data generated in social media platforms has unavoidably become an important source of understanding human opinions and behaviors. Models of sentiment analysis and tools to process social media data have been developed in English, German, Italian etc., but studies related to the Albanian language remain limited. The present paper aims to: i) provide a better understanding of the steps needed in a sentiment analyzer and, ii) present a model to automatically process tweets in Albanian language. The process starts with the cleaning and classification are the first steps of this process, while the generation of meaningful results is the final one. Entity and text analysis are used to provide the best insights and a better understanding of humans' opinion about different trending topics. The comparison between three machine learning classifiers (Naïve Bayes, SVM, Logistic Regression) is here made to address the best classification method. The performance of these classifiers is evaluated based on statistical accuracy tests.

Keywords**:** social media, classification methods, opinions

1. INTRODUCTION

Social media platforms are a key part of digital strategies of public and private institutions. Social media interaction is an umbrella term that encompasses all the two-way conversations and touchpoints that occur between the users. Twitter is one of the most used micro-blogging services nowadays, counting more than 330 million active users (Statista, 2021). The topology of the Twitter network is one directional, and each tweet is limited to 280 characters (Twitter Developer Platform, 2021). The use of Twitter data is linked to the provision of the Twitter API, that is a well-structured API which provides a high level of accessibility towards the information that users accept to make public.

Social media analytics involves the extraction, analysis and interpretation of the information generated on social media platforms, converting data into meaningful insights. The process of building an application for 'Social Media Mining' is reported in (Bonzanini 2016) as following:

- Authentication - linked with the OAuth (Open Authorization) standard used in Twitter;
- Data collection - the obtaining the information related to a user or topic;
- Data cleaning and pre-processing - the data need to be transformed in a readable format for further processes;
- Modelling and analysing - depends on the results that need to be achieved

Sentiment Analysis (SA) is a subfield of natural language processing which analyses the sentiment and emotions expressed by users related to a product/ service/ topic or any other user in the network (Rusell and Klassen, 2019). Observing sentiment from social media platforms is possible through the combination of NLP and machine learning tools. The SA techniques provide an automatic way to analyze conversations in social media, allowing organizations to learn their customer's opinion and needs. Different algorithms can be used in a sentiment analysis model such as: rule-based, automatic and hybrid (Rusell and Klassen 2019). In this study, we are going to use machine learning techniques (which are part of the automatic category) to learn from the data in disposal. While using these techniques we consider two main processes: training and prediction (Liu 2015). A training dataset is used to teach the model how particular inputs can be associated with an appropriate tag. After the model is trained, it can be used to predict the perfect tag that matches each input given in the model.

The present paper aims to find an appropriate solution to influence different social groups via data mining techniques. Python has been chosen as a programming language with a rich ecosystem, easy syntax and semantic which together provide the best means for both beginners and advanced users. Using Python and its associated libraries provides an answer to another important question: "How can all SA steps be established?". The main Python libraries used in each process are Python Docs, (2021):

- Tweepy: Twitter REST API is used to access previously generated tweets;
- NumPy (Numerical Python): To process effectively data structures in an array like form;

- SciKit-learn: To have access to the main classification algorithms that will be used in this work (Naïve Bayes, Logistic Regression and SVM);
- NLTK: To process and classify textual representations;
- Jupyter Notebook: To code live and directly view graphical representations and textual input.

The structure of this work is as following: Section 2 presents research studies linked with Sentiment Analysis; Section 3 explains the methodology followed; Section 4 presents the experimental results achieved and Section 5 provides the main conclusions achieved and possible future work.

2. Related Work

Many papers provide information about Sentiment Analysis for languages such as English, German, Chinese etc., but studies about Albanian language remain quite limited. The first approach for SA in Albania could be found in (Biba and Mane 2013). After experimenting with different classifiers, they concluded that the best performing classifier for their corpus was Hyper Pipes, with an average accuracy of 87%. Trandafili *et al.*, (2018) compared the performance of text classification algorithms in a corpus with 20 classes (40 document each). Using the Weka software to test different algorithms (Naïve Bayes, Random Forest, SVM, Decision Tree, K-Nearest Neighbor, Simple Logistic, ANN), they concluded that the best performing algorithm in their corpus was Naïve Bayes and SVM (Trandafili *et al*., 2018). Skenduli *et. al*., (2018) made a comparative analysis between different classifiers using the collection of around 60000 posts from Facebook, which belong to 119 Albanian politicians, and the results showed deep learning techniques having a better performance than other classical machine learning classifiers. Kote *et al.*, (2018) created 5 corpuses each containing 50 text documents of positive and 50 of negative opinions. WEKA software was used to perform experiments and evaluate the performance of classification algorithms. The result achieved after the experiments is that for different corpuses, different algorithms give different performances. Hyper Pipes was evaluated as the algorithm with the best performance (83.62%) in (Kote *et al*., 2018). Additional interesting information about text and emotion classification in Albanian language could be found in (Skënduli and Biba 2013; Voca and Kadriu 2015; Kadriu and Abazi 2017; Vasili *et al*., 2018; Kadriu *et al.*, 2019; Kote *et al*., 2019; Kote and Biba 2021; Vasili *et al*., 2021).

3. METHODOLOGY

Here, the steps needed to label posts in Twitter, assigning each post a positive or negative sentiment, is reported. The collection and processing of the available data is part of an elaboration process. Figure 1 depicts the model used, including all the steps needed for the tweet's classification. While trying to build a proper model for collecting and processing the available information, the necessary details to be taken in consideration are directly linked with the characteristics of the Albanian language and the format of each post. The forthcoming paragraph provides detailed information about the process of cleaning, processing, classifying and visualizing.

Main objectives

The present paper aims to; i) present different methods that can be used to develop a sentiment analysis model in Twitter and, ii) present the achieved results. It presents the necessary steps that can be followed to label each post with a sentiment, providing an overview of the main NLP methods for students and other interested individuals that are new in this field. The dataset created by (Mozetic *et al*., 2016) is used as a training and testing dataset. It contains datasets for 15 different languages, including Albanian language. The authors concluded that the quality of the training data directly affects the quality of the classification model (Mozetic *et al*., 2016). By using the dataset from (Mozetic *et al.*, 2016) as training and testing dataset, an opinion analysis could be carried out. To proceed with a concrete example, the analysis is going to be focused on extracting people's opinions and feelings from their posts in Twitter. The following steps are followed to achieve the results:

1. Pre-processing of the training data, so that they can be ready for the mining process (data cleaning, the removal of hashtags and other users mentioning);
2. Training and testing the classification algorithms with the (Mozetic *et al*., 2016);
3. The evaluation of the result from different classifiers

Using the Pareto Principle, 80% of the data is used for training purposes and 20% of the data is used for testing (Bonzanini 2016). To achieve the expected results in the classification module, we the Naïve Bayes, SVM, and the Logistic Regression classifiers were used. The classification algorithms are accessed from the NLTK Python library and are fed with the training data and then tested accordingly.

The proposed architecture

The architecture in the Figure 1 represents all the steps followed. The process of pre-processing and cleaning are described below.

In the end, the main result to be achieved by using this model is the automatic classification of the posts into two categories (positive and negative). In addition, an overview about the important aspects could be developed.

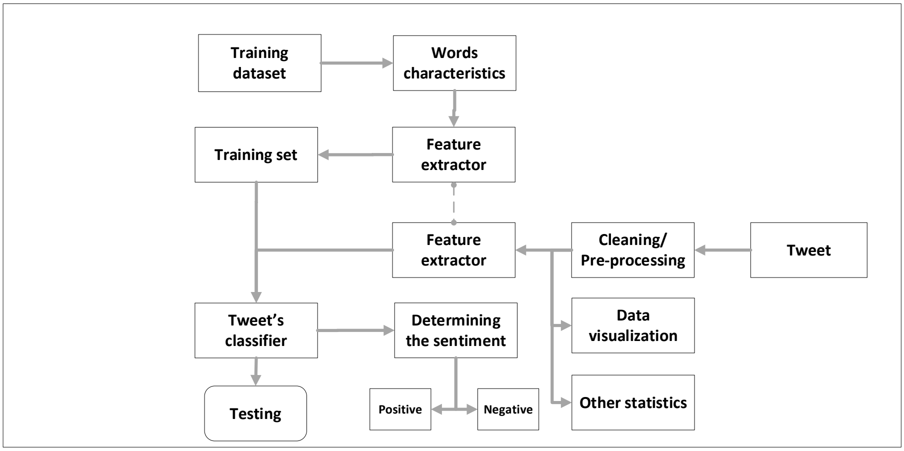

Fig. 7: The proposed architecture

Pre-processing

Having in disposal a previously created dataset (Mozetic *et al.*, 2016), it is important to pre-process the data to give them a proper format before the analysis steps. The pre-processing phase is one of the most important phases in the sentiment analysis journey, since it captures the most important words in a sentence or document. Different rules are determined depending on the language of the posts that we are collecting. In this phase we need to clean all the collected tweets, to find the proper message. Specifically, the steps that could be included in the pre-processing or cleaning phase are:

1. Transform all text into lowercase;
2. Remove stop-words; for Albanian language: ‘dhe’, ‘edhe’, ‘te’, ‘ne’, ‘sepse’, ‘por’ etc.;
3. Remove blank spaces, punctuations;
4. If necessary, remove words with less than three letters;
5. Remove hashtags (#), URL (http://....), users tagged (@);
6. Stemming- Consists in removing the end of the words;
7. Lemmatization- Removal of inflectional ending, return the base (lemma) of the word.

An example of cleaned tweet is in the Table 1 reported:

Table 4. Pre-processing a tweet

Original tweet	Në takim me @EPP President @JosephDaul: Moshapja e negociatave ndëshkon Shqipërinë dhe ndihmon politikanët që kanë tradhëtuar qytetarët shqiptarë, duke bashkëpunuar me krimin për pushtetin dhe pasurimin e tyre. #EPPSummit#Wearefamily#AlbaniaintheEU
Pre-processed tweet	takim, president, moshapja, negociatave, ndëshkon, Shqipërinë, ndihmon, politikanët, tradhëtuar, qytetarët, shqiptarë, bashkëpunuar, krimin, pushtetin, pasurimin

Table 1 shows that cleaning a tweet is an important process, as unnecessary elements are removed. Pre-processing provides a way to group the most important words in a post and offer an easier classification process.

The pre-processing function is built in Python and leads the way to the following steps. From the mentioned pre-processing steps, we have considered the following: transformation of the words in lowercase, removal of some of the most common stop words in Albanian, removal of blank spaces and punctuations, removal of URLs, hashtags (#) and user mentions (@). The stop words list is composed of a group of 173 stop words in total. The word_tokenize function from the NLTK's library is further used to return a tokenized version of the initial text, which in this point contains the main words that are linked with the text meaning.

Stemming and Lemmatization are not in the present investigation included. The use of proper stemming algorithm and lemmatization would further optimize the usage of this model, leaving space for future experiments and optimization.

Training and testing dataset

Mozetic *et al.*, (2016) said that dataset is used to train and test the classification algorithms. In addition, it is composed of around 53,005 posts, 8,106 of which express a negative sentiment, 18,768 neutral and 26,131 a positive sentiment. Since in this work only the positive and negative labels are taken into consideration, the total number of records is estimated to be 34,237. The dataset saves the following attributes for each tweet: Tweet_Id, Sentiment (Positive, Negative or Neutral) and the annotator's Id (Mozetic, 2016).

The majority of the techniques used to evaluate a model are rated according to the comparison between the label generated during the test and the proper label the input should have. The testing process has the same format as the training but the information differs.

Classification algorithms

To classify tweets in different classes (positive or negative) we need to build a classifier, which in this case is done with the help of the NLTK library. This library is one of the most powerful libraries in Python, at least for the classification algorithms included in it (Rusell and Klassen 2019). A script in Python is built to import these algorithms.

With the data in disposal in the training dataset, we can guarantee a satisfactory analysis of the Albanian posts, but there is a necessity to create a way for more detailed statistics with the help of a better constructed dataset in the future. The block diagram of this module is accordingly expressed in Figure 2.

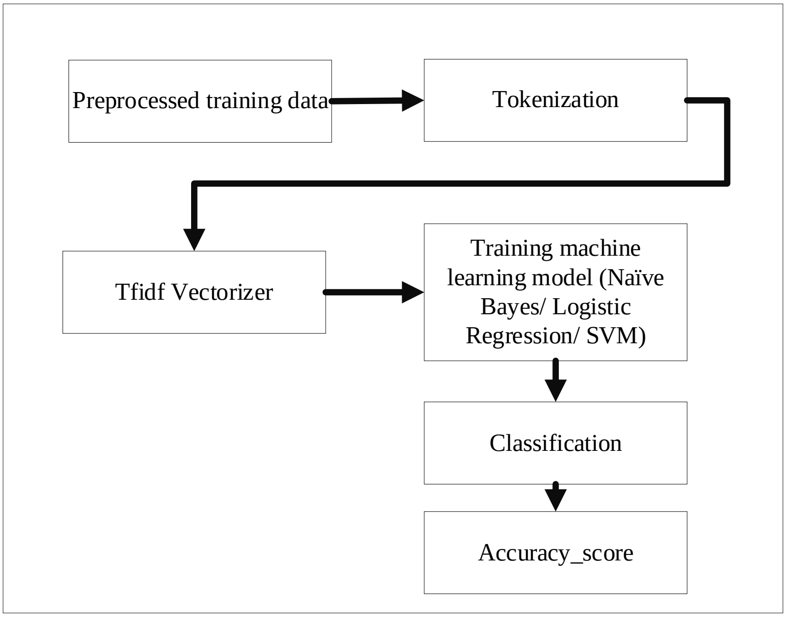
Preprocessed training data
Tokenization
Tfidf Vectorizer
Training machine learning model (Naïve Bayes/ Logistic Regression/ SVM)
Classification
Accuracy_score

Fig. 8: Block diagram of the feature extraction module

The list of all the expression contained in the respective dataset is transformed into a list of pairs (word, sentiment). The fit and predict methods from the sklearn library are used respectively in the training and testing dataset (PythonProgramming, 2021). The fit function calculated the mean and variance of each of the features in the data in disposal, while the predict function uses the calculated weights to make predictions in the test data. The used classifiers are Naïve Bayes, Logistic Regression and SVM.

The initialization of the classifier provides the possibility to test it with different collected tweets. The training process provided the classifier with the possibility to differentiate the presence of words and the sentiment related to them. In Jupyter Notebook we tried the following command:

```
tweet='Situata e sotme ishte e pakëndshme'
classifier.classify(extract_features(tweet.split()))
```

The classifying method (classify) takes as argument a group of tweet characteristics, which in this case contains the word 'pakëndshme' which is linked to a negative sentiment. The enlargement of the dataset would provide a better classification process. Once the training and testing process are performed, different data could be inputted in the model and produce a final targeting label.

4. Experimental results

The accuracy score function from the NLTK library is used to calculate the accuracy of each the classifiers. The evaluation of the classifier's accuracy is done by passing to the trained classifier the testing dataset. The results achieved are presented in Table 2.

Table 5. The accuracy of the classifiers

	Naïve Bayes	**Logistic Regression**	**SVM**
Accuracy	77.3%	79.7%	78.7%

Table 2 shows that the classifiers presented a good level of accuracy. The Logistic Regression classifiers present a higher accuracy rate for the dataset in disposal. For this reason, all the results achieved until now are relatively good to take into consideration. The classification and testing involved, provide a good model for the automatic classification of different posts regarding the sentiment they express.

By using this model, and the structure given, more advanced cases can be studied to achieve meaningful results for the future. We need to mention that by further checking the dataset and by using other pre-processing techniques, or even creating a more specific dataset, the classifiers accuracy in determining the sentiment may change and provide a more accurate final result.

5. CONCLUSIONS

This work provides the starting point towards sentiment analysis of text found in social media platforms. By providing all the steps that need to be followed the model represented is a good source of information for students and other interested individuals who are interesting in further developing their knowledge in natural language processing and opinion mining.

The built model can be used to determine a result for a specific topic or for a public person. The most important part is that by determining an Albanian language dataset, it is possible to classify the Albanian posts of different public personalities or Albanian companies. This model, even though is still in its first steps of implementation, can be further advanced to be helpful in different fields such as finance, marketing, innovation and so on.

Further on, a combination of the opinion and sentiment analyses linked with real time events, would provide a better possibility to understand what is happening around the world, or specifically in Albania. Political, artistic or sportive events can be processed in real time to gain a better insight about people's responsiveness.

Meanwhile, there are a lot of features to be fixed and updated so that to provide better results. From the review of literature and related works, we think that there is a necessity to create a well-structured public dataset in Albanian language for research and academic purposes. Further on, in this project only two sentiment groups were created (positive and negative). Adding other important categories such as neutral, angry, hope etc. would be beneficial in future cases. In the end, another important part, which would make the usage of this model and the part of visualization or classification more accessible, would be the development of a simple application where each person could search different topics or public personalities and find out the statistics related to them in graphical or statistical form.

REFERENCES

Biba M, Mane M. 2013. Sentiment Analysis through Machine Learning: An Experimental Evaluation for Albanian. *Advances in Intelligent Systems and Computing*.

Bonzanini M. 2016. *Masteing Social Media Mining with Python: Acquire and analyze data from all corners of the social web with Python*. Packt publishing.

Kadriu A, Abazi L. 2017. A comparison of algorithms for text classification of Albanian News Articles. *Entrenova*.

Kadriu A, Abazi L, Abazi H. 2019. Albanian Text Classification: Bag of Words Model and Word Analogies. *Business Systesm Research, 10*.

Kote N, Biba M. 2021. Evaluation of the Performance of Machine Learning Algorithms for Opinion Classification *International Journal of Innovative Science and Research Technology*, **6 (6)**.

Kote N, Biba M, Trandafili E. 2018. A thorough experimental evaluation of Algorithms for opinion mining in Albanian. *Advances in Internet, Data & Web Technologies*.

Kote N, Biba M, Kanerva J, Ronnqvist S, Ginter F. 2019. Morphological Tagging and Lemmatization of Albanian: A manually Annotated Corpus and Neural Models. ArXiv.

Liu B. 2015. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. doi:https://doi.org/10.1017/CBO9781139084789

Mozetic I, Grcar M, Smailovic J. 2016. Multilingual Twitter Sentiment Classification: The role of Human Annotators. doi:https://doi.org/10.1371/journal.pone.0155036

Python Docs. 2021. Retrieved 2021, from https://docs.python.org/3/library/

PythonProgramming. 2021. *Scikit-learn*. Retrieved from https://pythonprogramming.net/sklearn-scikit-learn-nltk-tutorial/

Rusell MA, Klassen M. 2019. *Miing the Social Web Data Mining Facebook, Twitter, LinkedIn, Instagram, Github, and more*. O'Reilly.

Skënduli MP, Biba M. 2013. A Named Entity Recognition Approach for Albanian. 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE.

Skenduli MP, Biba M, Loglisci C, Malerba D. 2018. User-Emotion Detection thorugh sentence-based classification using deep learning: A case-study with microblogs in Albanian. *International Symposium on Methodologies for Intelligent Systems, Limassol*, 258-267. doi:https://doi.org/10.1007/978-3-030-01851-1_25

Statista. 2021. Most popular social networks worldwide as of July 2021, ranked by number of active users, Retrieved 2021, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Trandafili E, Kote N, Biba M. 2018. Performance Evaluation of Text Categorization Algorithms Using an Albanian Corpus. *Advances in Internet, Data & Web Technologies*.

Twitter Developer Platform. 2021. *The structure of a tweet*. Retrieved 2021, from https://developer.twitter.com/en

Twitter Developers. 2021. *Rate Limiting*. Retrieved 2018, from https://developer.twitter.com/en/docs/basics/rate-limiting.html

Vasili R, Xhina E, Terpo D. 2021. Sentiment Analysis on Social Media for Albanian Language. *Open Access Library Journal*, **8**: 1-21, doi: 10.4236/oalib.1107514.

Vasili R, Xhina E, Ninka I, Souliotis T. 2018. A comparative Review of Text Mining & Related Technologies. *RTA-CSIT, Tirana*.

Voca B, Kadriu A. 2015. A scheme for Albanian Language Processing. DSC2015. Thessaloniki, Greece.