ISSN 2601-8683 (Print)
ISSN 2601-8675 (Online)
European Journal of
Formal Sciences and Engineering
July -December 2021
Volume 4, Issue 2

An Approach for Movie Review Classification in Turkish

Migena Ceyhan
Epoka University, Tirana, Albania
Zeynep Orhan
Union College, NY, USA
Dimitrios Karras
Epoka University, Tirana, Albania

Abstract

Web 2.0 has given to all people the right to become a representative of a huge cast of informal media. The importance of this power is getting more evident everyday. Every social media actor can influence the rest of the world by one's own opinions, feelings, and thoughts generously shared on multiple media. This information belonging to various fields of life can be very handy and be used to one's advantage, gaining precious experience. One of the greatest problems that this poses is the huge number of data spread everywhere, which are difficult to process as row data per se. Social media and general sentiment text analysis is of much valuable use, accomplishing the task extracting pure gold out of raw mineral. The key point of this investigation is to characterize new reviews automatically. To start with, features selected out of all the word roots appearing in the comments were used to train the system according to known machine learning algorithms. Next, critical words determining positive or negative sense were extracted. Another strategy was attempted eliminating common terms and dealing only with the significant class-determining words to build vocabulary with them. Aparts from linear approach, vector based feature sets were prepared out all or some of the features. The outcomes acquired were analyzed and compared leading to important conclusions, emphasizing the importance of feature selection in text classification.

Keywords: Sentiment Analysis, Turkish Language Processing, movie reviews, Machine Learning, Natural Language Processing

Introduction

The phenomena of Word of Mouth has always existed in any society, nevertheless its civilization degree, and it will always be present as long as we will be able to talk about

humankind. It only has changed means, shape, spreadability power, echo intensity, etc. Its power has changed a lot especially after the introduction of Web 2.0. Everyday strategies are renewed to match the dynamic content suggested by the global Word of Mouth. In the new context, any person decides for oneself to take part in the cast of a real life movie, deciding whether to just appear or to be a star. One does so by being an active and interactive participant in social life, which has no strict constraints. Social media provides people with such tools, that enable news reach in seconds any part of the globe, without even having the chance to verifiy them in advance. This is the power of the weapon of 21st century society.

It is estimated that out of 7.8 billion of the world population, more than 4.6 billion individuals (58% of the complete population) is dynamic in web, with North Americans having the most elevated entrance rate with around 95% of their population (Internet World Stats, 2020). As an illustration of a social media platform, Facebook gets visited at least daily by 74% of its clients in the US (Zephoria - Digital Marketing, 2020). People share their opinions and get quite influenced by others' opinions, too. This power of social media can also explain quite the common exploitation of it from the political and business leaders around the world. Each post is followed by countless comments, which reveal a lot about the support or opposition related to the post. Products are associated with several reviews from people who either suggest or complain about the product.

As an instance of a corporation that has appreciably advanced currently is Netflix. The wide variety of paying streaming subscribers differs as 21, 66 and 182 millions of customers respectively in 2011, 2015, and 2020. Only during the first six months of 2020, under the favorable effects of the pandemic situation, 26 millions of new subscribers have chosen this company (Statista, 2020). Before starting to watch for hours people don't mind querying for some moments others' opinions. Sometimes, however, this process can take more than minutes because of the abundancy of the reviews. Tools to summarize or automatically color the data according to their polarity, positiveness or negativeness, are crucial in this area. This is the main goal of this study, to predict movie comments in seconds. In the next sections, the following issues will be dealt with: similar works in this area, explaining the experimental setup, revealing and explaining the results, as well as including some conclusions regarding this study and potential future ones.

Literature Review

An extensive literature review conducted by Mostafa (2013) suggests that most Sentiment Analysis applications might be classified into four distinct categories: product reviews, movie reviews, politically oriented abstractions and stock market predictions. In this study overall efforts to analyze sentiments in online movie reviews, as well as some previous state of the art of the sentiment analysis of texts in Turkish will be covered.

In one study (Na, 2010) the opinion mining of movie reviews from discussion board threads, user reviews, critics' reviews and bloggers' postings were performed. Reviews sentence length, lexicon, and parts-of-speech information was considered, concluding that opinion holder words like verbs and adjectives were used more in reviews, respective to an abundance of nouns used in bloggers or critics posting. The most commonly used positive and negative words and their patterns per each domain were determined.

Zhuang (2006) applied ML techniques movie reviews dataset in order to summarize the opinion polarity. They aimed to retrieve specific feature sets in the text and the expressed opinion, for example, "sound effects" and "excellent". Pang (2002) similarly utilized support vector machines (SVM) acquiring 82.9% of precision in categorizing opinion summaries of movie reviews. They made use of both single words (unigrams) and pairs of consecutive words (bigrams) to classify the comments, as well of different classification algorithms, concluding that Naive Bayes, Maximum Entropy and SVM, have performed better in text classification than in sentiment classification. In fact, it is difficult to get better outcomes, because of specific characteristics of natural languages, but in specific domains, we can get good enough outcomes.

Thet, Na and Khoo (2008) conducted studies classifying online movie reviews. In one of them, they used machine learning and information extraction techniques by correctly determining the pronouns and co-referencing them to relate them to different aspects, such as the cast, the director, the effects, etc. and the overall rating of the movie. In the next task, the authors tried to reveal also the strength of the polarity of the comment according to a suggested computerized method, by taking into consideration the grammatical dependency structure of each clause analyzed according to a computational linguistics approach.

An advantage of the lexicon-based approach as compared to more generally used machine learning is that in the former, training set need not be labeled previous to the classification. They work according to text grammar analysis principles, while the elder fits the algorithms to the training set characteristic patterns. It is interesting that besides being inferior to machine learning methods in specific domains, lexicon-based methods can be quite better for wider domain sets. For instance, a lexicon-based approach (Taboada, 2011) is used for six distinct corpora from various domains, with a 75–80% accuracy. ML techniques, on the other hand, result more efficient for a distinctive domain, with 86.4% accuracy for a movie review summarization for a given dataset (Pang B. &., 2004). Once again it is implied that, although showing a weaker performance in data classification within one domain due to training dataset pattern overfitting, lexicon-based methods are more robust and show better results in cross-domain text classification process, in the current work, getting better scores in blog postings and video game reviews.

Sindhwani (2008) build a semi-supervised lexical model by merging lexical sentiment information, unlabeled data, and labeled training data. In three of the domains used in the study, such as products, political, and movie reviews, this strategy outdid purely supervised and competing semi-supervised approach.

There exist few studies in other natural languages, generally not applying very different methodologies than those used in text classification in English. Nevertheless, they are important in their field as novel applications in other languages. In a study in the Spanish language that has been conducted by Martínez-Cámara (2011) for movie reviews classification, they used several ML techniques and attained a high accuracy of 86.84% when SVM was applied.

Turkish is not an exception for languages other than English, where only a limited number of studies in sentiment analysis area exist. One of them (Kaya, 2013) studied the sentiment analysis of Turkish political columns on web documents. Their approach considered transfer learning in Turkish. In transfer learning, the aim is to extract needed knowledge from one or more tasks and then to transfer extracted information into a target task. In this work, the unigrams and the bigrams together with polar Turkish terms are used as classification characteristics to categorize unseen documents. The authors used four different classifiers: Naive Bayes, Maximum Entropy, SVM, and the character-based n-gram language model and compared their accuracies. They concluded that Maximum Entropy and n-gram language model is more efficient than SVM and Naive Bayes classifiers. The classification accuracy in different cases ranges from 65% to 77%. On the other side, several works have studied causal association rule mining.

Erogul (2009) investigated sentiment in two movie datasets in the Turkish language, applying English language sentiment analysis approaches. Turksent (2010) is an annotation tool developed specifically for manual sentiment analysis of Turkish social media posts. Yet another study in the Turkish language compares methods of text representation (Amasyalı, 2012). An ambitious study (Vural, 2013) aimed to determine the polarity of movie reviews by translating Sentistrength library to Turkish. They used a large corpus of Turkish movie reviews and they stated that although the framework was unsupervised, the performance approached the performance of supervised polarity classification. Amanet (2017) studied Twitter data using the emotion categories like “Happy", “Appreciation" etc., defining the most effective word sets for each emotion. Turkmen (2016) worked on the aspect-based sentiment polarity of online customer reviews. In Kamburoğlu (2018) thesis adjective clustering was used to automatically guess Turkish movie review scores of 76% accuracy. Through this study it was possible to measure also the reliability of the two popular sentiment lexicons SenticNet and SentiWordNet, resulting in a moderate level of agreement between lexicons and human judgments with an accuracy of 79%. Orhan (2014) automatically predicted the text polarity in customer product comments domain by making use of language characteristic features of the reviews and by utilizing ML techniques with a

high level of correctness. An optimistic upcoming aim of their research is to categorize texts on any topic.

Experimental Setup

Data selection

Movie comments from well-known and popular Turkish movie sites, such as (IMDb, 2014), (Sinemalar.com, 2014), and (beyazperde, 2014) were collected. Based on the high and low ranking of the comments, positive and negative were partitioned respectively. In total 305 positive and 305 negative comments were gathered, 205 of them used for training and 100 comments per group for both classes were used to test the system.

Data Preprocessing

Some preliminary tasks were done before selecting the feature sets, such as tokenization, stemming, morphological analysis and disambiguation, getting rid of some stop words, etc. The whole process is shown in detail in this section.

Turkish, together with Korean, Hungarian and Finnish belong to the Altaic language group. It has characteristics of vowel harmony and extensive agglutination, meaning the derivation of new words by adding derivational or inflectional morphemes to the roots as suffixes, resulting in many different words derived from one root or stem. These properties cause difficulties in NLP, because of the complex morphology compared with other languages like English. This makes the computerized processing of data analysis for Turkish more challenging.

Natural Language Processing (NLP) is one of the fields of science and engineering, useful to design computer systems for processing and understanding natural languages (Rehman, 2013). NLP is widely used also in computational linguistics, filling the gap between human expressions and artificial intelligence. The advances in information technologies have driven many studies dealing with natural languages. The basic NLP steps are tokenization, stemming, POS tagging, etc.

Tokenization is one of the preliminary steps of text processing. It is the process of separating sentence structure into word groups, which is applied in order to simplify the process of analysis extracting information from requirements documents (Webster, 1992). First, the sentences are separated into their components by making use of the punctuation marks and spaces as separators.

The next preliminary step is **stemming**. After the text has been tokenized into words, it is cleaned from the inflectional morphemes through the stemming process. Words derived from the roots of nouns and verbs through derivational morphemes, by removing the inflectional suffixes is named a stem (Can, Kocberber, Ocalan, & Vursavas, 2008). Derivational suffixes are used to derive words. The inflectional suffixes are added to the stem of the name and verbs to specify the state, possession, plurality, time.

However, stemming is a difficult task in agglutinative languages such as Turkish, Finnish, and Korean. Because the sequence of inflectional suffixes can be added to the stem of a word. The stem of the word is accessed by clearing the word from the inflectional by removing them.

Stemming has been done by using the **morphological analyzer** tool tr-tagger (Turkish Language Resources, 2019), a Turkish morphological tagger includes Kemal Oflazer's finite state machines for Turkish.

Next, to get rid of ambiguous tr-disamb, a Turkish **morphological disambiguation** tool was used (Turkish Language Resources, 2019). The tool shows a high accuracy of disambiguation of about 96%, based on the Turkish language rule model for morphological disambiguation. Turkish, being an agglutinative language, close to half of the words of a general text can show morphological ambiguity, which makes the disambiguation task quite tough. The correct form according to the context is chosen among an unlimited number of morphological analyses of a word due to multiple candidate suffixes.

Feature Selection

Feature selection is one of the most important tasks in Sentiment Analysis. After the initial processing of the data, the stems are acquired, which will form the features for different methods. All the unique stems emerging from positive and negative movie comments, together with their occurrences per document, and overall frequency per class, after some normalization procedure regarding the length of the text and the documents population, will contribute to positive and negative word lists. It is known that the most frequent words are usually common in both classes, so they cannot be distinctive. Examples of the common words can be pronouns, articles, etc. The words with similar occurrence frequencies, falling within specific thresh values will be eliminated either from both lists, as non discriminative. Then, different methods will be involved. Some of the concepts are explained as follows:

Thresh values: Some of the words appearing in similar ratio in both positive and negative word lists have been eliminated from both lists.

Erase methods: Some terms which fall between determined thresh values are erased from one or both lists, according to the erase method.

Erase from both lists method: The terms within thresh percentage similar frequencies are erased from both positive and negative lists

Erase from one list method: The terms within thresh percentage similar frequencies are erased from the list that they appear less often and their frequencies difference is kept in the other list

Binary List feature selection method: Initially all words (roots) are considered. According to which thresh ratio and erase method is going to be used, the positive words

list and negative words list are built. If the word used in the positive comment exists in the positive words list, then it contributes to the positivity of the comment as much as the frequency of the word used inside the same comment. The same is done for the negativity of the word from the positive comment, if the word appears in the negative word list for the given thresh and erase method, the frequency of appearance in the positive document is added to the negativity value of the comment. This is done for all the words of the comment. This process is repeated for all positive comments in the train set. The whole process is done also for all negative data from the train set. At the end of the process, a comment will have a positive and a negative value which will be the estimated class value (P for positive, N for negative determined from the difference of the negative value from the positive value calculated). Again all the above is repeated for the test data set, for positive and negative comments.

Frequency List feature selection method: The difference from Binary List method is that each word value occurrence in the document is multiplied by the frequency coming from the positive and negative words list of words. This calculation will affect the total positivity and negativity value of each comment. The rest of the process is as above.

All Words Binary method: All distinct words coming from train documents become the first row of the vector, while all the documents (id's or names) coming from the train dataset become the first column of the vector. If a word appears in the comment, its value is placed as 1, if it doesn't, its value equals 0. The last column contains the information for positivity or negativity of the comment. In the same way also the test data file is prepared, with the exact same first row. For words appearing in test dataset but not in train dataset, no value will be kept, obviously.

All Words Frequency feature selection method: The only difference from All Words Binary method is that not only the occurrence is determined, but also a numeric value is calculated regarding the number of occurrences of the word in the document, the frequency coming from the positive words list and the frequency coming from the negative words list.

Selected Words Binary feature selection method: Instead of taking all the words out of the train data test, according to the thresh value and erase method, a set of those words is taken as the feature set of the vector. The rest is done similarly to the All Words Binary feature selection method.

Selected Words Frequency feature selection method: Instead of taking all the words out of the train data test, according to the thresh value and erase method, a set of those words is taken as the first row of the matrix. The rest is done similarly to the All Words Frequency feature selection method.

This calculation will affect the total positivity and negativity value of each comment. The rest of the process is as above.

Classification with Machine Learning techniques

Machine learning based studies can be categorized into supervised, unsupervised and semi-supervised topics. Feature engineering and feature selection are also vital in a machine learning pipeline.

Supervised learning is one of the most used approaches in ML domain. Although supervised learning is successful in rich set of applications, it has many challenges

Lastly, the features for train and test data prepared by the above mentioned feature methods yield .arff extension files, which are fed to Weka Classification tool. Several existing Machine Learning algorithms are used and their results are evaluated. Voted Perception (VP), Bayesian Linear Regression (BLR), Random Forest (RF) and Logistic Regression (LOG). They can be seen in the Table 1:

Table 1: Accuracy values (%) for all feature selection methods and their ML classifiers

VP	BLR	RF	LOG	Feature Method
90	89	87	88	Frequency List EraseBoth 25% thresh
88	90	87	88	Frequency List EraseBoth 50% thresh
79	80	83	89	All Words Binary
85	89	84	79	All Words Frequency
86	88	83	72	Selected Words Binary EraseBoth 25% thresh
83	90	87	78	Selected Words Frequency EraseBoth 50% thresh

Results and Conclusions

When comparing Binary and Frequency List feature selection methods, Frequency List method clearly outdoes the Binary one to an extent 90% to 79% of respective accuracies.

It can be noted that the highest accuracies of 90% can be obtained from Frequency List feature selection method together when erasing the common terms from both positive and negative word lists within the thresh values 25% and 50% of the values, respectively for Voted Perception and Bayesian Logistic Regression ML algorithms. The same accuracy is obtained also with the Selected Words Frequency EraseBoth 50% feature selection method when classified with Bayesian Logistic Regression ML algorithms.

In general Bayesian Logistic Regression ML algorithm is very accurate, giving 89% value for Frequency List EraseBoth 25% thresh and All Words Frequency feature selection method. While Bayesian Linear Regression does not give a good result with All Words Binary feature selection method, Logistic Regression has an accuracy of 89%.

One can thus conclude that all feature selection methods have similar results with different ML algorithms. All the results are quite high for Turkish movie reviews when compared with previous studies in this field, making this one a promising study for Turkish and generally in this area.

Forthcoming research might exploit machine learning techniques for more significant feature refinement, reducing the feature sets in favor of performance and predicted accuracy.

References

[1] (2019, December). Retrieved from Turkish Language Resources: http://www.denizyuret.com/2006/11/turkish-resources.html
[2] Amanet, H. (2017). Sentiment analysis in turkish social media texts. Karadeniz Technical University.
[3] Amasyalı, M. F. (2012). Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması" - "A Comparison of Text Representation Methods for Turkish Text Classification. EMO Bilimsel Dergi, 2(4).
[4] beyazperde. (2014, May). Retrieved from http://www.beyazperde.com/
[5] Can, F., Kocberber, S., Ocalan, C. H., & Vursavas, O. M. (2008). Information retrieval on Turkish texts. (F. Can, S. Kocberber, C. H. Ocalan, & O. M. Vursavas, Eds.) Journal of the American Society for Information Science and Technology, 59(3), 407-421. doi:doi:10.1002/asi.20750
[6] Eroğul, U. (2009). Sentiment analysis in Turkish. Middle East Technical University, Ms Thesis, Computer Engineering.
[7] IMDb. (2014, May). Retrieved from https://www.imdb.com/
[8] Internet World Stats. (2020). Retrieved from https://www.internetworldstats.com/stats.htm
[9] Kanburoglu, A. B. (2018). Graph clustering approach to sentiment analysis. Işık University.
[10] Kaya, M. (2013). Sentiment analysis of Turkish political columns with transfer learning. Middle East Technical University.
[11] Martínez-Cámara E., M.-V. M.-L. (2011). Opinion Classification Techniques Applied to a Spanish Corpus. In: Muñoz R., Montoyo A., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2011. Lecture Notes in Computer Science. 6716. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-22327-3_17
[12] Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. Expert Syst. Appl, 40, pp. 4241-4251. doi:https://doi.org/10.1016/j.eswa.2013.01.019
[13] Na, J. T. (2010). Comparing sentiment expression in movie reviews from four online genres. Online Information Review, 34(2), 317-338. doi:10.1108/14684521011037016
[14] Orhan, Z. G. (2014). CUSTOMER SATISFACTION MEASUREMENT TOOL BY ANALYSING TURKISH PRODUCT REVIEWS. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 7(1), 12–18. https://dergipark.org.tr/en/pub/tb.

[15] Pang, B. &. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ArXiv, cs.CL. doi:0409058
[16] Pang, B. L. (2002). Proceedings Of The ACL-02 Conference On Empirical Methods In Natural Language Processing - EMNLP '02. doi:https://doi.org/10.3115/1118693.1118704
[17] Rehman, Z. A. (2013). Morpheme matching based text tokenization for a scarce resourced language. PloS one, 8(8), 1-8.
[18] Sindhwani V., M. P. (2008). Document-Word Co-regularization for Semi-supervised Sentiment Analysis. Eighth IEEE International Conference on Data Mining, (pp. 1025-1030). Pisa. doi:10.1109/ICDM.2008.113
[19] Sinemalar.com. (2014, May). Retrieved from www.sinemalar.com
[20] Statista. (2020). Retrieved from https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide/
[21] Taboada, M. B. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37, 267–307. doi:10.1162/COLI_a_00049
[22] Thet, T. N. (2008). Sentiment classification of movie reviews using multiple perspectives. Proceedings of the international conference on Asian digital libraries (ICADL), 184–193.
[23] Turkmen, H. (2016). Turkmen, H., “Discovering product features from Turkish reviews by using aspect based sentiment analysis. Kocaeli University.
[24] Turksent. (2010). Retrieved May 2017, from Annotation tool developed specifically for manual sentiment analysis of social media posts: http://www.turksent.com
[25] Vural, A. G. (2013). A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. Computer and Information Sciences III, 437-445.
[26] Webster, J. &. (1992). Tokenization as the initial phase in NLP., (pp. 1106-1110). doi:10.3115/992424.992434
[27] Zephoria - Digital Marketing. (2020). Retrieved from https://zephoria.com/top-15-valuable-facebook-statistics/
[28] Zhuang, L. J.-y. (2006). Movie Review Mining and Summarization. Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM).