EasyChair Preprint

№ 11571

Automatic Opinion Extraction from Football-Related Social Media: a Gazetteer and Rule-Based Approach

Atmane Hadji and Mohmed-Khireddine Kholladi

December 19, 2023

Automatic Opinion Extraction from Football-Related Social Media: A Gazetteer and Rule-Based Approach

Atmane Hadji
Department of Computer Science, University Center A . Boussouf Mila, 43000 Mila, Algeria .
a.hadji@centre-univ-mila.dz

Mohmed-Khireddine Kholladi
Department of Mathematics and Computer Science, HAMMA lakhdar , El oued University, El-Oued , Algeria. MISC Laboratory of Abdelhamid Mehri university of Constantine 2, Algeria.
kholladi@univ-eloued.dz

***Abstract*—Sentiment analysis on social networks has become a highly active area of research in recent years. With the explosion of social media and the massive amount of user-generated data, it has become crucial to understand the opinions and sentiments expressed online. Sentiment analysis is used to categorize expressed feelings in various ways, such as negative, positive, or neutral. The aim of this work is to enhance techniques for researching and extracting opinions. The main idea is to identify opinions within a set of documents or texts available online for exploitation by other systems. In this study, we present an approach based on an opinion detection system on social networks (Facebook) regarding the UEFA Champions League. The implementation of this solution was carried out using the GATE platform (General Architecture for Text Engineering). This work thus contributes to the field of sentiment and opinion analysis in social networks by employing Gazetteers and leveraging the JAPE rules (Java Annotation Patterns Engine).**

Keywords— Sentiment Analysis ; Social Networks; Gazetteer ;JAPE, UEFA Champions League.

I. INTRODUCTION

Thanks to the Internet, there is the possibility of discovering the opinions and feelings of a large number of people, which can be very useful in making informed decisions and forming an opinion on a given subject. In this work, we focus on analyzing opinions expressed regarding the UEFA Champions League. We study the information and comments posted on social networks (Facebook) in order to extract information about fan preferences, team performances, influential players, and emerging trends in the competition. Our goal is to understand the opinions of supporters and analyze the sentiments and preferences expressed towards the UEFA Champions League. Analysts and sports media use these opinions to assess the strengths and weaknesses of teams, predict match results, and provide comments and analyses to fans. People's opinions provide insight into the perception of the game, team strategies, and the impact of individual performances. In this context, we proposed an opinion analysis approach based on a Gazetteer. The aim of this work is to present some techniques to enhance the automatic detection of opinions and sentiments from comments on social networks. The goal of using a Gazetteer in opinion extraction is to strengthen the performance of the given results and improve their quality.

II. BACKGROUND AND RELATED WORKS

The extraction of opinions based on rules involves using predefined patterns or guidelines to identify and extract subjective information, sentiments, or attitudes expressed in text data. This approach is often used in natural language processing (NLP) and sentiment analysis tasks. Here's a background overview of opinion extraction based on rules:

- Rule-Based NLP: Rule-based NLP relies on a set of predefined linguistic patterns, grammatical rules, or heuristics to process and analyze text data. These rules are designed by linguists or NLP experts and are used to capture specific linguistic structures, sentiments, or entities within the text.
- Subjectivity and Sentiment Analysis: Opinion extraction is a subtask of sentiment analysis, which aims to determine the sentiment or emotion expressed in a piece of text. Subjectivity refers to the extent to which a statement is influenced by personal feelings, opinions, or beliefs.
- Key Components: Linguistic Patterns: Rules are typically defined based on linguistic patterns, syntactic structures, or semantic cues. These patterns may include specific keywords, parts of speech, or syntactic relationships that are indicative of opinions or sentiments.
- Gazetteers: A gazetteer is a list of words or phrases associated with specific categories or entities. It can be used in conjunction with rules to identify named entities or specific terms related to opinions.
- Regular Expressions: Regular expressions are powerful tools for defining complex patterns in text. They can be employed to capture various linguistic features that indicate opinions.

Our proposed approach is a hybrid approach merging between the two previous approaches cited.In the following, we will mention some related works :

The work proposes by [1] a rule-based algorithm for sentiment analysis, incorporating POS tagging during data preprocessing using Python. The user interface utilizes HTML with highlighted phrases, emojis, and GIFs to convey emotions. The analysis indicates that the proposed algorithm excels in predicting sentence sentiments and does so in a remarkably short time frame.

The authors [2] suggest a novel rule-based approach for extracting aspects from product reviews. This method relies on intuitive knowledge and the analysis of grammatical structures in sentences to identify both explicit and implicit aspects. The evaluation of the system on two popular datasets demonstrates that it achieves a higher detection accuracy compared to state-of-the-art aspect extraction techniques for both datasets.

The study realized by [3] investigates the application of aspect-based sentiment analysis in the legal domain to extract valuable information from legal opinion text. The authors introduce a rule-based approach for conducting aspect-based sentiment analysis, aiming to determine the sentiment expressed in a sentence regarding each legal party involved in a court case, considering these parties as the aspects of analysis.

III. APPROACH PROPOSED

The following architecture (Figure 1) depicts the detailed design of our opinion analysis system. The proposed system consists of several stages:

- **Data Collection** : We get information from social network (Facebook) online. We processed comments related to fan opinions semi-automatically.
- **Pretreatment** :In this step, we identified the comments related to the Champions League, then processed them in the next step.
- **Processing and Text Analysis**

We carried out the first phase of this work which is the research and collection of opinions on a sample of 70 comments from Facebook pages posted on the Champions League (chosen at random). after this phase we analyzed the text within the framework of automatic processing of a natural language by GATE [4] according to the order of steps: (Tokenization, Sentence Splitter, Part Of Speech Tagger, ...) .However, the gazzeeters were created for the indexing and extraction of opinions. The steps of text processing is as follows:

A. Tokenization

The Tokenizer divides text into simple words such as numbers, punctuation marks and many different types. For example, we have different words in Majestic and Minuscule, and among certain types of punctuation, etc. There is a "Token" annotation in the box, it should not be changed for different applications or text types.

Fig. 1. General architecture of the proposed system

B. Sentence Splitter

The sentence splitter is a cascade of finite-state transducers that segments text into sentences. This module is required for the tagger. The separator uses a list of gazetteer abbreviations to help distinguish phrase marking points from other types.

C. Part Of Speech Tagger

The tagger used is a modified version of the Brill tag, which assigns a part-of-speech tag to each word or symbol in the text. It is based on a lexicon and a set of default rules, which were learned from a large corpus from the Wall Street Journal. These elements can be adjusted manually if necessary.

Two additional lexicons are available: one for texts entirely in uppercase and the other for texts entirely in lowercase. To use them, simply load the appropriate lexicon, replacing the default one. In any case, the default rule set should always be used.

D. Semantic Tagger

The semantic marker of ANNIE (A Nearly-New Information Extraction System) is based on the JAPE (Java Annotation Patterns Engine) language. It uses rules that act on annotations, features and values assigned in previous sentences. It also includes rules that act on annotations, features and values that need to be assigned manually .

E. Gazetteer Creation

The Gazetteer plays a vital role in identifying named entities in the text based on predefined lists. It consists of different lists, such as city names, organization names, days of the week, etc. These lists are plain text files, where each entry is on a separate line. The Gazetteer contains not only specific entities, but also relevant keywords such as company abbreviations (e.g. "Ltd."), titles (e.g. "Dr."), etc. The lists are then compiled into finite state machine structures, which map text tokens to identified entities .

TABLE I. GAZETEER LIST EXAMPLE

	Positive Opinion	Negative Opinion	Neutral Opinion
Match Opinion	Exhilarating, highly entertaining, memorable match, Historic remnants Remnants, Historic match	Frustrating Underwhelming Subpar, Dismal, Mediocre Sloppy	neutral fan, evenly matched Balanced match, Thrilling game well fought , neutral observers
Team Opinion	Effective, Dominant Skilled,Cohesive, Effective Competitive	execution fell, vulnerable,underwhelming Weak,Inept,Inefficient	Evident,steady, Competent Capable, Decent, Adequate

In this study, we take two opinions, match opinion and team opinion, where supporters' opinions which are divided into a negative opinion, a positive opinion, and a neutral opinion, and the words which express which are among the lists from the gazetteers.

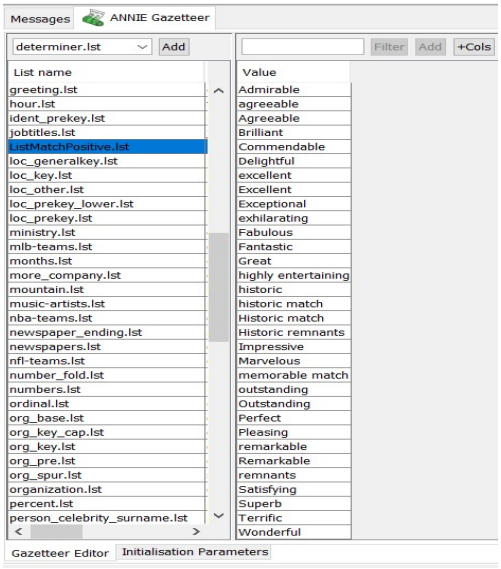

Fig. 2. Gazeteer list in GATE of Opinion Match Positive and Nigative

F. JAPE Rules :

A JAPE grammar [5] consists of a set of phases, each of which consists of a set of pattern/action rules. The phases execute sequentially and constitute a cascade of finite-state transducers on the annotations. The left side (LHS) of the rules consists of a description of the annotation pattern. The right side (RHS) consists of annotation manipulation instructions. Corresponding annotations on a ruler's LHS can be referenced on the RHS by means of labels attached to the pattern elements. Below is an example of a JAPE rule.

```
phase: championsleague
Input: Token Lookup
Options: control = appelt

//****** Opinion Audience Positive ******//
Rule:OpinionAudiencePositive
(
  {Token.string=="enthusiasm"}
  {Token.string=="ignited"}
  {Token.string=="the"}
  {Token.string=="stadium"}|

  {Token.string=="audience"}
  {Token.string=="created"}
  {Token.string=="an"}
  {Token.string=="incredible"}
  {Token.string=="atmosphere"}|

  {Token.string=="fans"}
  {Token.string=="were"}
  {Token.string=="passionate"}
  {Token.string=="passionate"}|

 {Lookup.majorType == "location"}

):adel
-->
:adel.OpinionAudiencePositive ={Rule="OpinionAudiencePositive"}
```

Fig. 3. Example of JAPE Rules

IV. RESULTS AND EVALUATION

After running the corpus with the use of JAPE and Gazetteer rules (figure 4), the system is now able to detect the entities named "Opinion Match Positive", "Opinion Match Negative" and "Opinion Match Neutral" corresponding to opinions on a Champions League match.

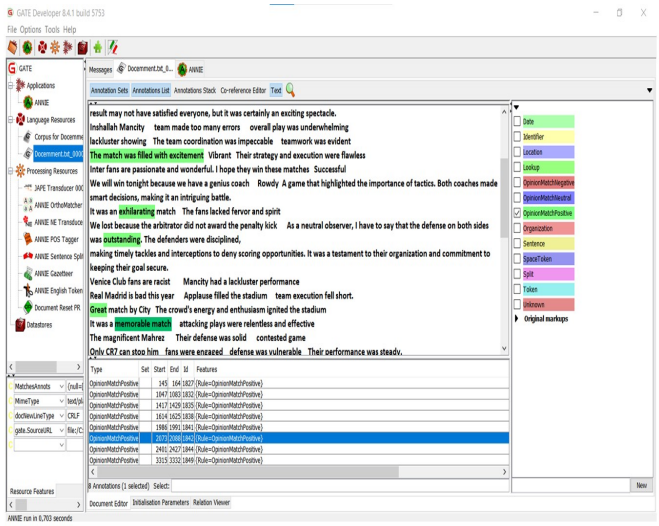

Fig. 4. Gazeteer list of Opinion Match Positive and Nigative

We processed a corpus containing 70 comments of 1000 words from social media (Facebook) using JAPE rules and using the Gazeteer we calculated the percentage: Recall,

precision and F-measure are widely used measures in NLP assessments;

- Precision is the percentage of correct results among the results obtained .

$$\text{Precision} = \frac{\text{Number of NE correctly recognized}}{\text{Number of NE recognized}}$$

- Recall is the percentage of correct results among the results that must be found .

We present formulas for evaluation such as precision, recall which are widely used measures in NLP evaluations

$$\text{Recall} = \frac{\text{Number of correctly recognized NE}}{\text{Number of corpus NE}}$$

- The F-measure is the combination of precision and recall and their weighting. The formula for the F-measure is as follows:

$$\text{F} - \text{mesure} = \frac{2(\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

V. Analysis and Discussion

A. Information Analysis

According to the (Table II) it shows that opinions are extracted from two main categories: “Match” and “Team”. Each of these categories is then subdivided into positive, negative and neutral opinions.

When it comes to opinions on the matches, there is a relatively balanced distribution between positive (20), negative (16) and neutral (17) opinions. This suggests that match comments are quite varied in terms of the sentiments expressed.

Regarding opinions on teams, we notice that positive opinions (29) are more frequent than negative (16) and neutral (9). This could indicate a more positive trend towards teams.

TABLE II. Opinion Extraction Results

	Named Entity		Precision	Recall	F-mesure
Positive Opinion Match	20	38	0.71	0.80	0.75
Negative Opinion Match	16	30	0.65	0.84	0.73
Neutral Opinion Match	17	32	0.67	0.84	0.77
Psitive Opinion Team	29	54	0.80	0.85	0.82
Negative Opinion Team	16	29	0.62	0.69	0.65
Neutral Opinion Team	9	17	0.70	0.76	0.73

The results obtained in this study are highly satisfactory, as demonstrated by the good Precision and excellent Recall rates (see to Table II and Figure 5).

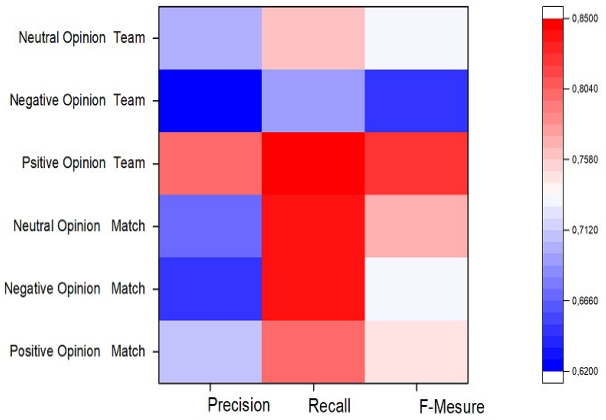

Fig. 5. Our Approach Evaluation

B. Overall Performance:

Overall, the performance of named entity opinion extraction appears to be quite robust, with acceptable precision, recall, and F-measure values for most categories. These results indicate that the system is capable of correctly identifying opinions in the text.

C. Category-wise Analysis:

The results are divided into three main categories: "Positive Opinion", "Negative Opinion", and "Neutral Opinion", each further subdivided into "Match" and "Team". Let's examine them in more detail:

- **Positive Opinion Match:** This category exhibits high precision (0.71), indicating that the majority of mentions extracted as positive opinions are indeed accurate. The recall is also strong (0.8), demonstrating that most positive opinions have been captured. The balanced F-measure (0.75) suggests a good trade-off between precision and recall.
- **Negative Opinion Match:** Performance is also good for this category, with a precision of 0.65 and a recall of 0.84. The F-measure of 0.73 indicates a good balance between precision and recall.
- **Neutral Opinion Match:** This category shows similar performance to the previous categories, with a precision of 0.67, a recall of 0.84, and an F-measure of 0.77.
- **Positive Opinion Team:** This category displays high precision (0.8), solid recall (0.85), and a balanced F-measure (0.82), indicating very good performance in extracting positive opinions about teams.

- **Negative Opinion Team:** While performance is lower than that of the "Positive Opinion Team" category, it remains acceptable, with a precision of 0.62, a recall of 0.69, and an F-measure of 0.65.
- **Neutral Opinion Team:** This category exhibits relatively low precision (0.05), indicating that the majority of mentions extracted as neutral opinions are incorrect. However, the recall is high (0.76), showing that most neutral opinions have been identified. The F-measure of 0.73 reflects a balance between precision and recall.

D. Comparison of Categories:

Overall, the "Positive Opinion" and "Negative Opinion" categories have similar performances, with balanced precision and recall. The "Neutral Opinion" categories have lower precision but higher recalls. The results demonstrate that extracting positive and negative opinions is more successful than extracting neutral opinions, especially for teams. It might be useful to explore specific strategies to improve the precision of neutral opinions.

In summary, the results show that the system is capable of extracting opinions with generally good performance. However, there are variations across categories, suggesting potential for continuous improvement, particularly in extracting neutral opinions. These results provide a solid foundation for opinion analysis in the context of matches and teams. They can be valuable for sports analysts, media, and sports organizations to evaluate public reactions and trends.

VI. CONCLUSION

After In conclusion, our study is part of the context of the analysis of sentiments and opinions in social networks, with a particular emphasis on the Football Champions League. Through the use of advanced techniques such as Gazetteers and JAPE Rules, we have developed an innovative approach to extracting and understanding opinions expressed online. Our work demonstrated the effectiveness of this approach, allowing more precise detection and better quality of the extracted opinions.

Leveraging the GATE platform, we have implemented an opinion detection system on Facebook, providing a valuable tool for analysts, sports media and football fans. This methodology can be extended to other domains and social networks, thus opening new perspectives for sentiment analysis research .

References

[1] U. Vudatha ,K. Prema '' Aspect Based Sentiment Analysis Using Rule Based Approach '' First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT) , India, 2021.

[2] P. Soujanya , C. Erik , K. Lun-Wei , G. Chen , G. Alexander ''A Rule-Based Approach to Aspect Extraction from Product Reviews'' Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 28–37, Dublin, Ireland, 2014.

[3] R. Isanka , R.M. Chanika , K. Dilini ,S. Nisansa S ,R. Gathika , A. Shehan Pererak ,'' Rule-Based Approach for Party-Based Sentiment Analysis in Legal Opinion Texts ''https://arxiv.org/pdf/2011.05675.pdf

[4] GATE: http://gate.ac.uk/, last accessed 2022/12/20.

[5] D. Thakker, P. Photosand ,R. Tanaka, Taha Osman, P. Lakin,,(2009.) ''GATE JAPE Grammar Tutorial Version 1.0. UK: PA Photos'',

[6] Gate.ac.uk ," Chapter 8 JAPE : Regular Expressions over Annotations,"[Online],Available: https://gate.ac.uk/sale/tao/splitch8.html.

[7] P. Malo., A. Sinha , P. Takala , P. Korhonen ,J. Wallenius , "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts", JASIST, 2014.

[8] LI F., '' The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach''. Journal of Accounting Research, vol. 48, no 5, p. 1049-1102, 2010.

[9] T. Loughran , B. McDonald , "The Use of Word Lists in Textual Analysis", Journal of Behavioral Finance, vol. 16, no 1, p. 1-11,2015.

[10] X. Mao , S. Chang , J Shi ,F. Li , R. Shi , "Sentiment-Aware Word Embedding for Emotion Classification", Applied Sciences, vol. 9, p. 14,2019.

[11] P. Kumar ,J. Prabhu , "Role of sentiment classification in sentiment analysis: a survey", AISC,vol. 1168, p. 196-209, 2018.