UDK: 316.6:796.332(100)“2018“
Prethodno priopćenje
29. XI. 2021.

DINO AVDIĆ*
MARINA BAGIĆ BABAC**

APPLICATION OF AFFECTIVE LEXICONS IN SPORTS TEXT MINING: A CASE STUDY OF FIFA WORLD CUP 2018

ABSTRACT

World Cup is a major football event that is globally popular and has its very best influence on human emotions. As such, it affects how people verbally discuss football topics on the Internet. In addition, it shows great significance when viewers who usually do not watch other football competitions start paying close attention when their nation plays a World Cup football match. In this paper, fans' online behaviour during World Cup 2018 was analysed using text mining methods. With the use of emotion analysis, it is noticed that there are different emotional states through which people go while sharing their thoughts with other people about football. Reddit, a discussion Internet website, was used as a generator of user data. Five supervised machine learning algorithms were used to test and revise an existing model. It is affirmed that the model successfully predicts the emotions within the text with an average accuracy of 78%.

Keywords: text mining, emotion analysis, football, Reddit, machine learning.

* Faculty of Electrical Engineering and Computing, University of Zagreb, dino.avdic@fer.hr
** Faculty of Electrical Engineering and Computing, University of Zagreb, marina.bagic@fer.hr

1. INTRODUCTION

People have always been involved in sports-related conversations. Sport has thus become an integral part of many people's lives. Before the emergence of the Internet, people had such discussions mostly in smaller human groups. Today, an inhabitant of one continent can discuss with another inhabitant of another continent without needing to be with him in person. It is due to the constant development of the Internet as a fast communication medium. With this speed of Internet development, a huge amount of information has been collected by people, making it easier for them to connect, especially on sports topics.

Football is, as it is often said, the most important secondary thing in the world and that is what gives more interest to this research. It is the most popular sport in the world with an estimated 3.5 billion fans.[1] According to FIFA, the total number of unique viewers of the 2018 FIFA World Cup exceeded 3 billion across all viewing methods, with more than 1 billion people projected to have watched the final game. World Cup 2018 is on track to be the most-viewed sporting event ever on digital platforms. When it comes to attention, the significance of the World Cup has surpassed even the significance of the Olympic Games. Therefore, there is no better laboratory to examine emotional reactions than football games during World Cup.

This paper studies the emotional states through which people went during football events of the World Cup 2018 focusing on the text of football discussions using emotion mining and affective lexicons. The testing of proposed models is done using several machine learning algorithms (Caruana and Niculescu-Mizil, 2006), which will predict the quantity of each emotion in comments. The results obtained in this paper can help a deeper understanding of the emotional states and relationships between people concerning sports. The first part of the paper refers to the data - process of data collecting, processing, and modelling for the analysis. The second part of the paper deals with the results of the analysis.

1 https://www.fifa.com/worldcup/

2. RELATED WORK

Sporting events have long been serving as a natural laboratory to study cognitive processes and emotions (Gilovich, Vallone, and Tversky, 1985). Sporting events provide a rich source of data against which to test psychological theories (Bestgen, 2008) because they follow uniform rules, which are repeated many times, and elicit a variety of measurable emotions (Russell, 1980). There are a lot of theses that deal with emotion analysis in sports, but there are only a few with the analysis of the emotions focused on the football area. One that uses emotion analysis in the football-related texts (Gratch, 2014) is mentioned below. Others mentioned below are also similar theses related to sports emotions analysis specifically (Russell, 2003) and emotion analysis generally (Russell, 1980).

It is generally accepted amongst players, coaches, and fans that momentum is a major factor in determining outcomes in sporting events (Tversky, 1985). Early success by one side within a game is presumed to create confidence, positive emotions, and continuing success, whereas early failure creates negative emotions and continuing failure. The effect that is known as the hot hand fallacy (Gilovich, Vallone, and Tversky, 1985) can be applied to a wide range of sports.

A study by Fernández-Dols and Ruiz-Belda (1995) examined the facial expressions of gold medal winners in the Olympic games. It was trying to distinguish theories that argue that emotional expressions reflect true feelings from competing theories that represent those facial expressions as communicative acts. Based on manual coding of facial displays, the winning athletes don't smile when they learn they have won but only upon turning towards the audience. They concluded that expressions serve as a social signal rather than a marker of emotion.

Another study (Gilovich, Madey, and Medvec, V., 1995) used sporting events to lend support for prospect theory, one of the most important theories in behavioural economics. Prospect theory predicts that people's feelings towards an outcome depend on their reference point (i.e., the same event may be seen either as a loss or a gain depending on one's point of comparison). While

examining the emotional reactions of players during the 1992 Olympics, this theory found support in the fact that silver medalists felt worse (because they felt the loss of not winning) than bronze medalists (who felt relieved for making it to the podium).

Sugimoto (2004) addressed sentence-level emotion recognition with a model that uses a composition assumption: the emotion of a sentence is a function of the emotional affinity of the words in the sentence. They obtain emotional judgments of 73 adjectives and a set of sentences from 15 human subjects and compute words' emotional strength based on the ratio of times a word or a sentence was judged to fall into a particular emotion bucket, given the number of human subjects. In addition, Alm, Roth, and Sproat (2005) classified the emotional affinity of sentences in the narrative domain of children's fairy tales for subsequent usage in the appropriate expressive rendering of text-to-speech *synpaper*.

Furthermore, Read (2005) used emoticons in newsgroup articles to extract instances relevant for training polarity classifiers, while Mishne (2005) and Yang, Hsin-Yih Lin, and Chen (2007) used emoticons as tags to train SVM classifiers at document or sentence level. In their studies, emoticons were taken as moods or emotion tags, and textual keywords were taken as features.

Another study (Clore and Ortony, 2008) used the emotional reactions of basketball fans to provide support for their appraisal theory of emotion. Throughout an entire college basketball season, they asked 106 fans to record their appraisals and emotions before, during, and after all 20 games in one year's season. The data provided clear support for their structural theory and their model of emotional intensity.

Gratch, Lucas, Malandrakis, Szablowski, Fessler, and Nichols (2014) explored the theory of emotions using sentiment analysis to examine tweets posted during World Cup 2014. They reviewed the importance of sporting events as a natural laboratory to study human behaviour, especially emotion. They described a corpus collected from the 680 million tweets. They used sentiment analysis techniques over this dataset to examine a theory from sports economics concerning what makes a sporting event exciting.

Staiano and Guerini (2014) made a vocabulary of emotions on about 37 thousand words that had an emotional score (from rappler.com). When it comes to creating an affective lexicon, they have demonstrated the benefits of collecting data from social networks. Similarly, Bandhakavi, Wiratunga, Massie, and Padmanabhan (2017) made generating lexicon through emotional detection from the text. Various datasets such as blogs, news, Twitter were used for the classification of words by emotion. Using another approach, Tang, Wei, Qin, Liu, and Zhou (2014) extended the traditional word embedding methods (Mikolov, Sutskever, Chen, Corado, and Dean, 2013; Collobert, Weston, Bottou, Karlen, Kavukcuoglu, and Kuksa., 2011) by encoding sentiment information into the existing continuous representation of words. They built sentiment-specific word embedding by developing three neural networks wherein the sentiment polarity of the tweet is incorporated in the neural networks' loss functions.

Mohammad, Bravo-Marquez, Salameh, and Kiritchenko (2018) presented the SemEval-2018 Task 1: Affect in Tweets, which includes an array of sub-tasks where automatic systems must infer the affectual state of a person from their tweet. Some of the tasks are on the intensities of four basic emotions common to many proposals of basic emotions, namely anger, fear, joy, and sadness. Some of the tasks are on valence or sentiment intensity. They included an emotion classification task over eleven emotions commonly expressed in tweets, which were determined through pilot annotations. Furthermore, Mohammad and Kiritchenko (2018) created a large single textual dataset annotated for many emotion dimensions (from both the basic emotion model and the VAD model). Specifically, they annotated tweets for the emotions of people that posted the tweets—emotions that can be inferred solely from the text of the tweet. For each emotion dimension, they annotated the data for not just coarse classes (such as anger or no anger) but also for fine-grained real-valued scores indicating the intensity of emotion (anger, sadness, valence, etc.). They showed that certain pairs of emotions are often present together in tweets. For example, the presence of anger is strongly associated with the presence of disgust, the presence of optimism is strongly associated with the presence

of joy, etc. Moreover, Mohammad (2019) presented the NRC VAD Lexicon, which has human ratings of valence, arousal, and dominance for more than 20,000 English words. He used Best–Worst Scaling to obtain fine-grained scores and address issues of annotation consistency that plague traditional rating scale methods of annotation. He showed that there exist statistically significant differences in the shared understanding of valence, arousal, and dominance across demographic variables such as age, gender, and personality.

3. RESEARCH METHOD

For this study, the required data were collected from the Reddit website, the largest Internet discussion platform. Four pre-processing phases were conducted to get words from comments. Then, the lexicon-based and supervised approaches were applied to extract the emotion of each comment. Finally, the obtained results were analysed.

The dataset is obtained programmatically by using Python Reddit API Wrapper. Discussions are divided into threads, and each thread represents a topic. For discussions on soccer (football) topics, a thread called "soccer" has been selected. The comment replies are also included as the study's purpose is to analyse the emotions and opinions of many people on the Internet. The total number of collected comments is 152,585. The dates of extracted comments origin are in the range from 14th of June to 15th of July, which is the duration of World Cup 2018.

For the data processing, the programming language R is used. It is selected because it supports many statistical functions, is fast in working with large data sets, and is suitable for displaying statistical analysis results.

For every comment, to extract words from them, four steps were conducted as below:

- Splitting text into words
- Converting words to the lowercases
- Removing stopwords
- Removing punctuations

In addition, all meaningless words were removed. Also, for every word two additional values are stored in the data table: a number of occurrences (frequency) and a number of characters in it (length). The average frequency of a word is 14 and the average word length is around 9.

	word	Freq	len
1	like	11169	4
2	just	11049	4
3	can	8880	3
4	game	8683	4
5	world	8439	5
6	team	7698	4
7	one	7241	3
8	cup	6556	3
9	don	6529	3
10	goal	6357	4

Figure 1. Top 10 frequent words

4. DATA ANALYSIS AND RESULTS

In this section, an emotion-based approach to text analysis is applied. Then, to evaluate how supervised machine learning can be applied for emotion analysis of the comments (Bagić Babac and Podobnik, 2016), we started with assessing which classification algorithm provides better accuracy. The emotion score of Reddit comments is analysed using the *syuzhet* package developed in the R programming language for the emotion and sentiment analysis. A method for extracting emotion, which uses the NRC sentiment dictionary developed by Mohammad (2018), was used. The comments are classified into eight emotion groups of “anger”, “anticipation”, “disgust”, “fear”, “joy”, “sadness”, “surprise”, “trust”. Most comments consist of mixed emotions, involving more than one emotion recognised by analysis. Analysing our dataset, 23% of comments were classified as anger, 27% as anticipation, 17% as disgust, 22% as fear, 21% as joy, 20% as sadness, 15% as a surprise, 27% as trust. In Table 1, a few examples of actual comments are illustrated.

Table 1. Example of emotion evaluation on the Reddit comments

Comment	Emotion score	Class
It's a symbol of Albania. Represents the double-headed eagle on their flag. It's the hand signal both Xhaka and Shaqiri made after scoring.	7	anger
If Croatia win I hope Modric gets paraded around the pitch as if he himself is the trophy (which he is).	7	anticipation
Damn all those world-class players in the offense for Argentina and they can't score shit. With such a dirty display they deserve this plus Croatia was the better side. Hurts to say as a Chelsea fan as well but Willy is a very bad gk (love the man tho)	5	disgust
Holy shit. My mom came into my room to bring me a plate of chicken nuggets and I literally screamed at her and hit the plate of chicken nuggets out of her hand. She started yelling and swearing at me and I slammed the door on her. I'm so distressed right now I don't know what to do. I didn't mean to do that to my mom but I'm literally in shock from the results tonight. I feel like I'm going to explode. Why the fucking fuck did Durmaz foul? This can't be happening. I'm having a fucking breakdown. I don't want to believe the world is so corrupt. I want a future to believe in. I want Zlatan to be president and fix this broken team. I cannot fucking deal with this right now. It wasn't supposed to be like this I thought that Kroos was shit???? This is so fucked.	6	fear
To be honest in hindsight the first Swedish goal is a blessing to the Germans it's only after that goal Sweden started to have more confidence played more open and toe to toe with the Germans ... which led to many opportunities. It's almost felt like the Germans are both lucky and unlucky at the same time in this game b/c they failed to capitalize on soooo many good chances and yet the football God decides to let them score a impossible goal in the very last seconds ...	8	joy
Hard to believe any black or gay football fan would risk going to this World Cup knowing Russia's history.	4	sadness
Eng winning in pens - bloody hell!! Nothing less would do in this wonderful unpredictable World Cup. What a roller coaster its been from the very 1st match and we aren't even in the quarters yet. Best WC ever. And since it is not been said enough - wonderful hosts. Screw all the garbage agenda-driven UK/US media who said Russia would be a disaster and warned visitors.	5	surprise
I really have to question the tactics Low threw out there today more than anything. Consistently crossing the ball which makes sense with the height advantage but you'd think after X amount you'd reconsider your approach to the game. On top of that saving the majority of your subs till the end of the match when you are 1-0 down kind of blows my mind. Maybe he was trying to keep team cohesion by leaving everyone on the field but it was painfully obvious that their forward moment was terrible. Werner was not good but you only bring Gomez on in the 80 something minute. Germany is skilled no doubt but Low got it wrong today and a lot of the players did not live up to their skill level. You have to wonder what the German camp was thinking during this game.	10	trust

Figures 2-7 show how the emotions were distributed during the whole period of World Cup 2018, which were the dominant ones and in which quantity was measured. Five stages of the tournament (with an illustration for the whole WC period) are covered here: the group stage, the round of 16, quarterfinals, semi-finals, and the finals.

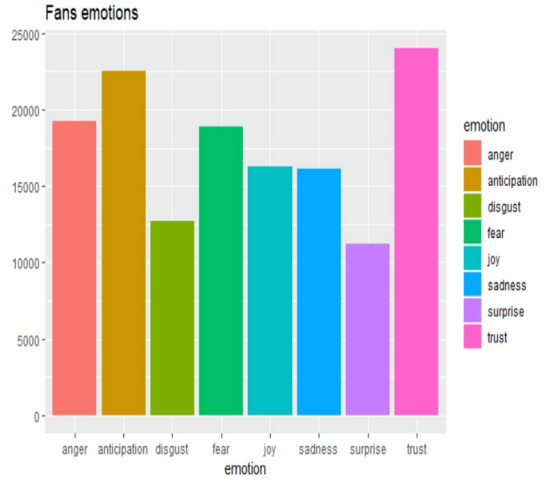

Figure 2. Emotions during group stage

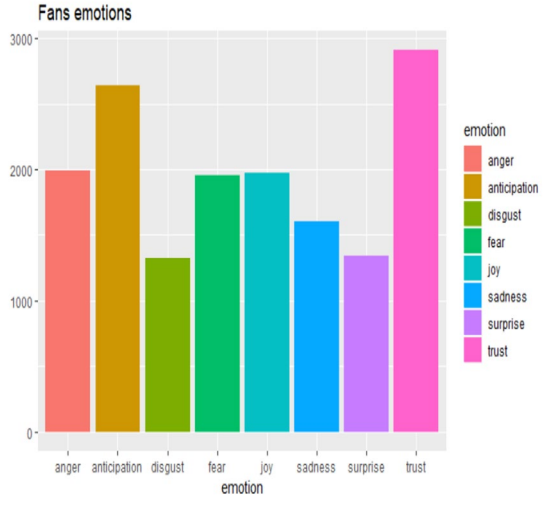

Figure 4. Emotions during quarterfinals

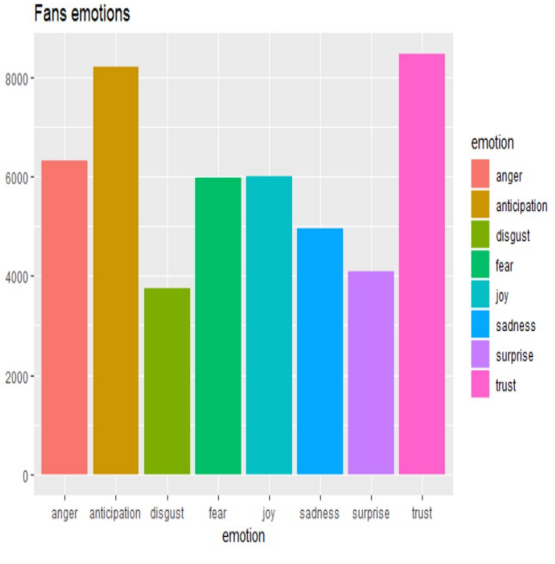

Figure 3. Emotions during the round of 16

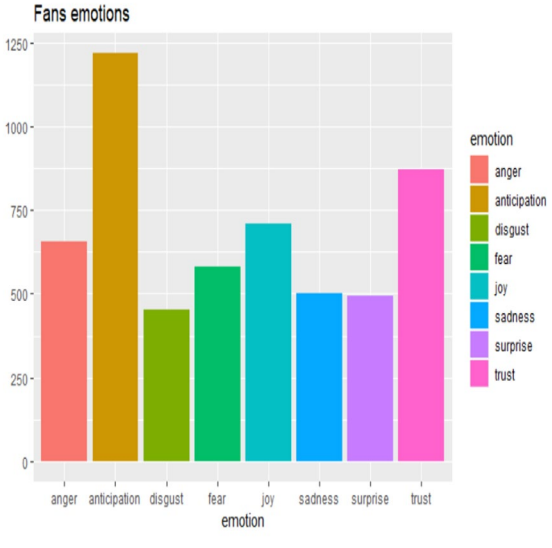

Figure 5. Emotions during semifinals

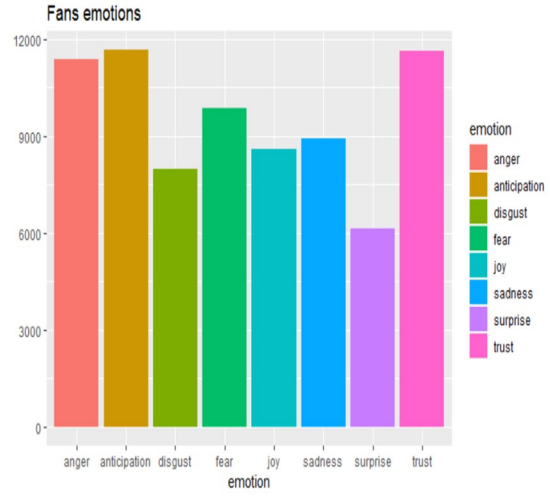

Figure 6. Emotions during the final game

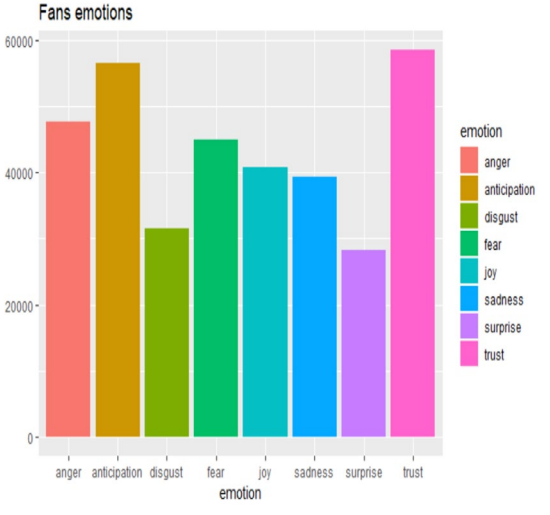

Figure 7. Emotions during the whole WC 2018

The above charts show two dominant emotions, anticipation and trust, during the World Cup phases. On the other hand, other emotions were also meaningful for each stage. For the group stage, anger and fear stood out while joy and sadness were following each other with almost the same quantity. For the round of 16 and quarterfinals, the charts are the same, as the group stage chart stand out with anger and fear. Also, in this case, joy is a little more present. The chart showing semi-finals emotions has anticipation as an emotion that highly stands out. Also, fear is less present than in the previous stages. In the final game, quantities of all the emotions increased due to a larger number of comments and the ending of the major sports event World Cup.

Next, intending to find an optimal algorithm to predict the comment sentiment, we used the datasets obtained based on the lexicon-based approach as training data. Then, the classification algorithms were applied for learning the outcomes of the emotion class of comments. The results were validated and tested based on their accuracy. Next, the stated data set was divided into train and test (90% for training, 10% for testing) to find the best algorithm (Zliobaite, 2015).

Five machine learning algorithms were used for measuring the accuracy of the generated emotion scores, namely, Naïve Bayes, Support Vector Machines, Neural Networks, K-Nearest Neighbor, and Decision Trees.

Naive Bayes is a supervised machine-learning algorithm based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature (Zhang, 2004). An advantage of Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. Despite their naive design and oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the implausible efficacy of Naive Bayes classifiers (Zhang, 2004). Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests (Caruana and Niculescu-Mizil, 2006).

Support Vector Machines is a supervised machine-learning algorithm used for classification and regression tasks that originated from statistical learning theory. As a classification method, SVM is a global classification model that generates non-overlapping partitions and usually em-

ploys all attributes. The goal of SVM is to identify an optimal separating hyperplane that maximizes the margin between different classes of the training data. The hyperplane which optimally separates different classes is the one that correctly classifies all the data while being farthest away from data points (Caruana, and Niculescu-Mizil, 2006).

Neural networks are a set of algorithms that are designed to recognize patterns. They help to group unlabelled data according to similarities among the example inputs, and they classify data when they have a labelled dataset to train on. The most basic type of neural net is something called a feedforward neural network, in which information travels in only one direction from input to output. These neural networks possess greater learning abilities and are widely employed for more complex tasks such as learning handwriting or language recognition (Tang, Wei, Qin, Liu, and Zhou, 2014).

The K-nearest neighbours is a supervised machine-learning algorithm used for classification and regression. It works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbours. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. It is called lazy because it doesn't learn a discriminative function from the training data but memorises the training dataset instead. A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example because they tend to be common among the *k* nearest neighbours due to their large number (Coomans et al., 1982). One way to overcome this problem is to weigh the classification, considering the distance from the test point to each of its *k* nearest neighbours. The class (or value, in regression problems) of each of the *k* nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a self-organising (SOM), each node is a representative (a centre) of a cluster of similar points, regardless of their density in the original training data. *K*-NN can then be applied to the SOM.

Decision Trees is a supervised machine-learning algorithm in which the data are continuously split according to a certain parameter. The tree can be explained by two entities, namely, decision nodes, and leaves. Each internal node denotes a test on an attribute, each branch denotes the test outcome, and each leaf node holds a class label. A decision tree is one of the fastest ways to identify the most significant variables and the relation between two or more variables (Poch Alonso, and Bagić Babac, 2022). Decision Trees are commonly used in operations research and operations management. When in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as the best choice model or online selection model algorithm. Another use of Decision Trees is for descriptive means when calculating conditional probabilities.

Table 2. Results for the performance of machine learning algorithms (in percentages)

	Naive Bayes	SVM	Neural networks	KNN	Decision tree
anger	76,3	78,1	76,9	78,0	76,3
anticipation	73,1	74,8	74,4	75,0	74,5
disgust	82,5	82,5	83,5	84,0	84,5
fear	76,6	77,2	77,8	78,0	78,0
joy	78,4	79,9	79,0	79,0	78,1
sadness	79,0	80,1	80,0	80,0	80,5
surprise	84,0	84,3	85,1	84,0	85,3
trust	71,4	72,6	71,2	72,0	72,3
AVERAGE	77,7	78,7	78,5	78,8	78,7

All algorithms approximately gave the same results, around 78%. It was revealed that KNN was a little more accurate in predicting than SVM and Decision tree, which are showing an accuracy of around 78,8%. After SVM and Decision tree, which gave the same average accuracy of 78,7%, neural networks gave the average accuracy of

78,5%. Finally, Naive Bayes showed the lowest accuracy with an average accuracy of 77,7%. Table 2 displays the results of the classifiers' validation separately together with the accuracy criterion (percentage of correct predictions).

CONCLUSION

Across the world, there are dozens of football fans who support their nations during the World Cup. World Cup is a major football event that is globally popular and has its very best influence on human emotions. It affects how people verbally discuss football topics on the Internet (Bagić Babac and Podobnik, 2016). Also, it shows great significance when viewers who usually do not watch other football competitions start paying close attention when their nation plays a World Cup football match. Therefore, in this paper, human behaviour was analysed observed from the texts during World Cup 2018.

It is noticed that there are different emotional states through which people pass while writing down thoughts related to football topics. An existing model was revised and tested. It successfully predicts the quantity of each emotion within the text, with an average accuracy of 78%. The marketing experts can use the study results when designing content on social networks. They can also serve a variety of football-related experts to analyse and track fans and their opinions, ideas, and expectations.

The paper results open future possibilities for discussion related to emotion analysis in sports (Poch Alonso and Bagić Babac, 2022). This study also gives a lot of answers to questions about relationships between people when it comes to football. Besides statistical, the study can be overviewed from the psychological and sociological perspectives.

REFERENCES

Alm, C. O., Roth, D., Sproat, R. (2005). Emotions from the text: Machine learning for text-based emotion prediction. *Proceedings of the Joint Conference on HLT–EMNLP.* 1-8.

Bagić Babac, M., Podobnik, V. (2016). A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football", Online Information Review, 40 (6), 814-833. DOI: https://doi.org/10.1108/OIR-02-2016-0050

Bandhakavi, A., Wiratunga, N., Massie, S., Padmanabhan, D. (2017) Lexicon Generation for Emotion Detection from Text. *IEEE Intelligent Systems*. 1-7.

Bestgen, Y. (2008). Building Affective Lexicons from Specific Corpora for Automatic Sentiment Analysis. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 1-5.

Caruana, R., Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proc. 23rd International Conference on Machine Learning*. 161-168.

Clore, G. L., & Ortony, A. (2008). Appraisal theories: How cognition shapes affect into emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), Handbook of emotions (pp. 628–642). The Guilford Press.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2493–2537.

Coomans, D., Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, 136, 15-27.

Fernández-Dols, J., Ruiz-Belda, M. A. (1995). Expression of Emotion Versus Expressions of Emotions. *Everyday Conceptions of Emotion, 81*, 1-13.

Gilovich, T., Vallone, R., Tversky, A. (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive psychology,* 17, 295-314.

Gilovich, T., Madey, S.F., Medvec, V. (1995). When less is more: Counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality and Social Psychology*, 69, 603–610.

Gratch, J., Lucas, G., Malandrakis, N., Szablowski, E., Fessler, E., Nichols, J. (2014). GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion. 2015 *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1-6.

Mikolov, T., Sutskever, I., Chen, K., Corado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems.* 3111-3119.

Mishne, G. (2005). Experiments with mood classification in Blog posts. *Style - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR,* 1-8.

Mohammad, S. (2019). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1-11.

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in tweets. *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018).* 1-17.

Mohammad, S., Kiritchenko, S. (2018). Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018).* 1-12.

Poch Alonso, R., Bagić Babac, M. (2022) Machine learning approach to predicting a basketball game outcome, *International Journal of Data Science* (in press)

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, 43-48.

Russell, J.A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology,* 39(6), 1161-1178.

Staiano, J., Guerini, M. (2014). DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 2, 1-7.

Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).* 208-212.

Sugimoto, F. (2004). A method to classify emotional expressions of text and synthesize speech. *First International Symposium on Control, Communications and Signal Processing,* 611-614, doi: 10.1109/ISCCSP.2004.1296469.

Yang, C., Hsin-Yih Lin, K. Chen, H. (2007). Building Emotion Lexicon from Weblog Corpora. *Proceedings of the ACL 2007 Demo and Poster Sessions*. 1-4.

Zhang, H. (2004). The Optimality of Naïve Bayes. *FLAIRS2004 conference*. 1-6.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv*. 1-21.

PRIMJENA AFEKTIVNIH LEKSIKONA U ANALIZI SPORTSKIH TEKSTOVA NA PRIMJERU SVJETSKOGA NOGOMETNOG PRVENSTVA 2018.

SAŽETAK

Svjetsko nogometno prvenstvo najveći je globalno popularni nogometni događaj koji ima veliki utjecaj na ljudske emocije, kao i na način izražavanja na temu nogometa na internetu. Toliko je važan da gledatelji koji obično ne prate ostala nogometna natjecanja pridaju veliku pozornost kada njihova nacija igra utakmicu na Svjetskome prvenstvu. Stoga je u ovome radu obrađeno ponašanje ljudi za vrijeme Svjetskoga nogometnog prvenstva 2018. godine metodama analize teksta. Analizom emocija utvrđeno je da postoje različita emocionalna stanja kroz koja ljudi prolaze dok razmjenjuju svoja mišljenja s drugim ljudima o nogometu. U ovome se radu za dohvaćanje korisničkih podataka koristio *Reddit*, internetska platforma za raspravu. Primijenjeno je pet algoritama strojnoga učenja kako bi se testirao i revidirao postojeći model te je utvrđeno da postojeći model predviđa emocije unutar teksta s prosječnom točnošću od 78 %.

Ključne riječi: analiza teksta, analiza emocija, nogomet, Reddit, strojno učenje.