

# MOBILE CROWDSOURCING NETWORKS FOR SENTIMENT ANALYSIS CHALLENGES

Mohammad Jafarabad<sup>1</sup> and Rouhollah Dianat<sup>2</sup>

<sup>1</sup>P.hD student, Department of information technology Engineering, Qom University, Qom, Iran  
tcsms@yahoo.com

<sup>2</sup> Faculty member, Department of information technology Engineering, Qom University, Qom, Iran  
rouhollah.dianat@gmail.com

**Abstract - One Of The Most Important Parameters Of Text Mining Is The Accuracy And Efficiency Of Performance. In Some Cases, Data Mining And In Many Cases Text Mining, When Facing The Challenge Of Precision, We Use Active Learning. In Active Learning, Part Of The Machine-Related Process Is Accomplished By Crowdsourcing. Often, Especially In Sentiment Analysis Projects, There Is The Need To Use Crowdsourcing Activity Under The Text Mining Subcategory. In This Article, Not Only Do We Introduce A More Detailed Crowdsourcing And Sentiment Analysis But Also Using A Mobile Network Tool Operating Crowdsourcing With A New Proposal Has Been Suggested. This Article Has Been Able To Increase The Incentive For Crowdsourcing By Defining Digital Currency And Replacing It With Digital Goods And Ultimately Increases The Percentage Of Mobile Network Usage For Crowdsourcing.**

**Keywords -** Mobile Network, Crowd Sourcing, Sentiment Analysis, Mining

## 1. INTRODUCTION

In recent years, natural language processing has shifted to research such as SEMEVAL and annotation. Recently, due to the emergence of platforms such as AMT and CF (Crowd Flower) and simultaneous increase in the percentage of Internet usage, crowdsourcing has become one of the fastest and cheapest methods for text processing and annotation research. Crowdsourcing is used to do jobs in mechanized way, where workers get paid for their duties. Crowdsourcing is used extensively in sentiment analysis. An example of these activities is data labeling, where users label the data according to their sense. Some crowdsourcing projects are conducted in several languages.

## 2. FROM DEFINITION PROCESS TO THE CROWDSOURCING PROJECTS PERFORMANCE

In task designing, simplicity must be taken into account as much as possible. Providing a comprehensive instruction helps this simplicity. Research in 2018 showed that there should be no more than 10 average questions asked from the audience, which is often reduced in tasks classification [1]. In large documents, it should be considered whether to assign the whole to a single task or to divide it into several tasks? Specifying the exact time and type of reward will have a significant impact on the quality and timing of data collection. Payments are always paid when the job has been verified and there is a possibility of no payment if cheating is detected. [2]

If a significant discrepancy is found, the reason will be checked manually in order to see whether there is cheating. This should be done by the verification performance software. Another way is to put 20% of the gold standard results in the task and control the answers with this 20%. The workers' opinion must be close to the truth. Challenge setting is also effective. For instance, dedicating a special award for the first person can be effective in terms of the quality or the timing of the tasks. [3]

Setting up an experimental trial helps to correctly identify the time and average cost of each task. Between \$ 0.05 and \$ 0.01 is mostly paid for trivial jobs. Research has shown that increase in perk, increases the number of workers but does not have much effect on quality [4].

For data preparation, a user interface must be designed. Designing the user interface can be a complex and large task. It is possible to extract corpus automatically with the help of user interface and perform language processing models. Iterative Development Research on language processing is a good infrastructure in this field. For example, pos labeling is done in this section. This user interface helps projects management and also monitors progress and evaluation better.

The AMT interface has attempted to detect cheating as much as possible. [5] Users cannot paste texts into this interface and have to type project-related texts manually, which prevents copying and pasting results from the Internet. The AMT interface shows the progress of the task and also analyzes it. This analysis helps to increase accuracy in research. If a preliminary run fails, we will change the condition and run it again. In [6] \$ 100 was paid for 10 days of trial run. Supporting an iterative approach will lead to continuous improvement. Even low

and high paying projects have the possibility to perform for a few minutes. There are other challenges along the way, such as worker training, screening and how to deal with workers' argument. However, the most important challenge is how to motivate and retain workers for other tasks.

One of the vital issues in such projects is the way in which the recruitment is done. In some cases it is done through job advertisements. These advertisements can be through blogs, websites, forums and social networks. In order to attract more workers, some have proposed combining several small projects. Some projects require that workers should be screened. For example, only a few workers are selected from each region. Some screening approaches include an emphasis on expertise. Some screening methods include local employment, performance based employment or qualification based employment.

The overall quality of the data received should be assessed to ensure the accuracy of the results. Tracking employees' performance over time, including checking the amount of work done, can be a criterion to measure quality. Majority vote or some computational algorithms can help us figure out the rate of quality. F-measure, Recall and Precision are also some of the research indicators of quality assessment.[7]

Privacy must be respected in projects. The lack of precise instructions for implementation can lead to the loss of participants' rights. In order to cooperate with workers, they should not be deceived and exploited. Low wages (under \$ 2 an hour) are also morally wrong. If good policies are designed to do the job, the project can be sustainable. People involved in this fundraising project are looking for sharing their knowledge in achieving scientific results, and using results that are contrary to ethical issues is not correct. Minimizing human supervision is one of the most main goals in integrated projects. The process of performing an integrated task for a word processing task is presented in Table 1.

### 3. Investigation of research in the field of sentiment analysis

Receiving public opinions about social events, marketing activities and product prioritization has caught the attention of the academic community and the business world. Meanwhile, the fields of opinion mining and sentiment analysis are combined. Although these two terms are often used interchangeably and even referred to as synonymous, opinion mining is the extraction and analysis of human beings' views of an entity, while the sentiment analysis focuses on identifying the emotions expressed in the text and then analyzing them.

Table 1. Steps in implementing a crowdsourcing of a text mining project

Type of Activity	Performing Explanations
NLP Problem Selection and Crowdsourcing Style	Often active learning in crowdsourcing style is chosen because it is used a lot in language processing [21]. Simplification, semantic similarity, and sentiment analysis are also commonly used in language processing research. Some sentiment analysis tasks in the field of text mining require crowdsourcing.
What the crowdsourcing Worker Has to Do.	The user does not perform any particular language processing activity and just comments on data mining, which is often labeling or annotating. All language processing is done on the machine side. Of course, it is also possible to create a software and make it available to the crowdsourcing worker.
Collection and Pre-processing of the Body	There are ready corpus or webs. Pre-processing will include separating the root and eliminating repetitive and fluent words.
Guidance for Crowdsourcing Workers [8]	A software training course (Crowdsourcing Website) will be prepared as a pamphlet. If AMT is used, this collection will be available for them through the site.
Preliminary Implementation of Studies	Preliminary implementation and testing will be done on a small number with very low payments often based on dollars or euros.

Sentiment analysis is not a novel and newfound concept. It has been a long time in investment, business, and marketing using newspapers, journals and such, are studying consumers' opinions so they can see the influence of a particular company, brand or product.

Sentiment analysis is a field of study that analyzes the views, feelings, evaluations, assessments and attitudes of people and their feelings towards an entity such as: products, services, organizations, individuals, issues, events, topics, and characteristics. In general, the purpose of sentiment analysis is to determine the attitude of a spokesperson or author with regard to some topic or overall pole of a document. The classification of opinion on a subject, generally consists of positive, negative and neutral opinions. It is assumed that in the text, the comment is made on one topic. Nowadays, the best way is to measure the distance of the object between the positive or negative poles.

In the first stage, the intended object is located close to the word that gives us more information about the sense. The second stage involves identifying the semantic orientation obtained by calculating the known polarity of words. The third stage involves calculating the mean semantic distance from the pole for the keywords pair. In studies, between 71 and 85% of polarity has been correctly diagnosed [9].

One of the ideas mentioned in emotion analysis is document level. The work done at this level is to classify the views on the document. For instance, a review is received from the product, and the system detects whether the review expresses a positive or negative opinion about the product. There are many ideas about the sentence level problem. At this level of analysis the sentences are examined and it is anticipated that the belief expressed in the sentence is positive, negative or neutral (usually neutral means no opinion) [10, 11]. The present study examines polarity at the document level, a news item.

The main purpose of sentiment analysis is to determine the polarity of the text [12]. Polarity of the text means if the feeling of the text is positive or negative. Polarity determination is done at different levels. Polarity can be specified at document level, sentence level, characteristics level and entity level. At this level, the documents or sentences expressed are analyzed to determine their feelings. Machine training techniques are used to determine polarity. By using classification algorithms such as support vector machines, decision tree, maximum entropy and so on, it is possible to perform the necessary operations and the emotion in the text can be extracted [13, 14].

Classification is one of the data mining techniques that maps data to the predicted class and group. This technique provides intelligent decision making and not only is used to study existing samples, but also predicts the future behavior of the same sample. Classification consists of two phases: first the training phase, in which the data set is analyzed and in the second phase the data is tested and the accuracy of the classification pattern is obtained [15].

Text classification, which means assigning text documents based on content to one or more predefined classes, is one of the most important issues in text mining; sorting e-mails or files in a hierarchical order of folders, identifying the subject of the text, searching structure and / or finding documents that are in the user's interest, are among applications of text classification (classification, classification) [16]. Next step for classification is to extract the appropriate attributes among the document's expressions. This step is a critical step and the accuracy of the system is highly dependent on the selected keys that represent the document. Then, by using existing algorithms, the classification is conducted.

Considering the above, analyzing the emotions, classifying and identifying the tendencies or inclinations of the users on different topics are important. For this reason, this topic has been of interest to researchers. They seek to select a classification that can predict the polarity of data closer to the original author's point of view. Table 2 presents some recent research in the field of textual sentiment analysis with the aim of evaluating the identity of the text.

Table 2: Recent research in the field of sentiment analysis

Article	Definitions and Challenges	Function
[17]	The semantic orientation of specific features of the document has been studied in sentiment analysis. In corpus-based approaches, the value of the vocabulary pole is determined by the frequency of occurrence. Whereas in dictionary-based methods, the distance from the traditional Wordnet vocabulary is used. Eventually, the polarity of the document is calculated by calculating the semantic orientation. The analysis of specific features expresses the direction of the view.	This article has also been able to incorporate rare features in sentiment analysis. The problem with many methods of analyzing the feeling is being unable to take advantage of the combination of different ontologies, which this article has solved by presenting a plan.
[2]	In this article the role of public opinion in social events and marketing strategies has been examined. Keeping the structure of the information received on the Internet up to date is difficult. In this article, a study on detection of digressive semantic information has been done.	This study uses the TSVD algorithm to process cognitive information.
[18]	A great number of studies have been done in analyzing media discourse. In this article, the advances, which are the results of easier access to information by natural language processing, for automatic analysis of emotions in Western media	As there are about 600 reporters from 322 news agencies in China, this country is looking to find out which reporters report balanced news. On the other hand, examining the type of global attitude toward

	<p>news reports have been used. The steps include identifying linguistic bases, retrieving expressions from the news and analyzing emotions in news texts. Information sources are collected online and then analyzed using text mining techniques. Due to China's role in international trade, the world has definitely focused attention on developments in the country. This article examines sentiment analysis in China-related news in Western media.</p>	<p>China helps this country to refine its policies. To do this, a data recovery framework is required according to the following system.</p> <p>In this system, we first assume that we are going to examine the type of view of a particular word. First, the Chinese word is translated into the English equivalent. These steps include web crawling, bilingual searching, and statistical analysis of phrase similarity.</p> <p>2) Gathering a knowledge base that includes ontology. Wordnet and Hownet can be used to derive relationships.</p> <p>3) Recovering objective facts: Recovering the truth of the news, as it happened, is one hundred percent impossible. In general, news is developing around a main point. The rest of the story contains details. The main point is always to indicate the position of the concerned reporter or news agency. News might be defined for specific purposes. It is also possible that the journalist select a piece of news and continue the rest as he wishes with story-telling type, to convince the reader of his own goals. So choosing a news item from different news agencies may represent paradoxical facts. Therefore, at this stage, it is necessary to select corpus from news agencies with different perspectives.</p> <p>4) The fourth step is the analysis of emotions; sentiment analysis is one of the most difficult classification problems in language engineering. In this section the distinction between derogatory and commendatory words in terms of meaning as well as the difference between the intensity and the voice of vocabularies. The strategy that words can help to identify this feeling.</p>
<p>[19]</p>	<p>This article examines news sentiment analysis using crowdsourcing. This article has an approach to extract the scores of 37,000 words. This approach is implemented by the rappler.com website. This article uses an unsupervised learning approach. Many methods of emotion analysis use lexical sources, in most articles they look at the word out of context and say that each word has a positive and negative polarity. Most efforts are aimed at broad data coverage and high scoring accuracy. In some studies, some rules have also been developed to analyze emotions in sentences. Semantic similarity, word bags, and machine learning techniques are the most important methods used in sentiment analysis.</p>	<p>Rappler is a social news network that intelligent conversations, storytelling and interactive activities with the community are done there. People's feelings about the stories were extracted from the Rappler site. An MDE Document Emotions Matrix was constructed that polled eight feelings in each document. Feelings included fear, confusion, anger, annoyance, triviality, happiness, comfort, and sadness. In this article, a feeling is extracted for each document rather than a separate score for each sense. Their documents on this site were ranked 32% above all. This can be due to the cultural characteristics of the audience, the site's policy of spreading positive news, or the psychological phenomenon that people are interested in. The sense of happiness has also been explored in other articles.</p>

[20]	In this article there is a classification to detect emotions illustrated on Twitter. The researchers used the idea behind the NRC Canada system and implemented changes such as adding features and adding glossaries to their system using sparse SVM (l1-regularized). The system uses n-grammatical word components instead of n g features. They found that the n gram word components were as useful as n grams.	In this article on average, there are 843 features for per tweet. The team was among the top 10 participants at SemEval 2018 by earning an F1-score of 54/67 on Tweets.
------	---	---

#### 4. Offering a suggestion for mobile based crowdsourcing

If a suggestion is proposed for crowdsourcing, it is important to understand how much network capacity has been used to implement the offer. [127] Most crowd sources use the Internet, but it is also vital how many use the capacity of mobile networks instead of computers. In the proposed plan for crowdsourcing, using the meaning of mining and Bit coin simulation we were able to employ a large population using mobile for crowdsourcing.

In the proposed plan we performed the extraction process similar to Bit coin extraction and placed the extraction of digital currency instead of extraction of the goods. That is, each person received a percentage of a real good after the crowdsourcing process. The cheapest product in Iran is about \$ 0.014, and we consider 0.1 for each extraction. Due to the low value of Iranian money, this model was implemented at a low cost. In the first 3 days, 13,000 times the crowdsourcing process was completed and 1300 products were extracted. The cost of running this process was \$ 18, with each process costing \$ 0.013, which is lower than the global average of 0.02.

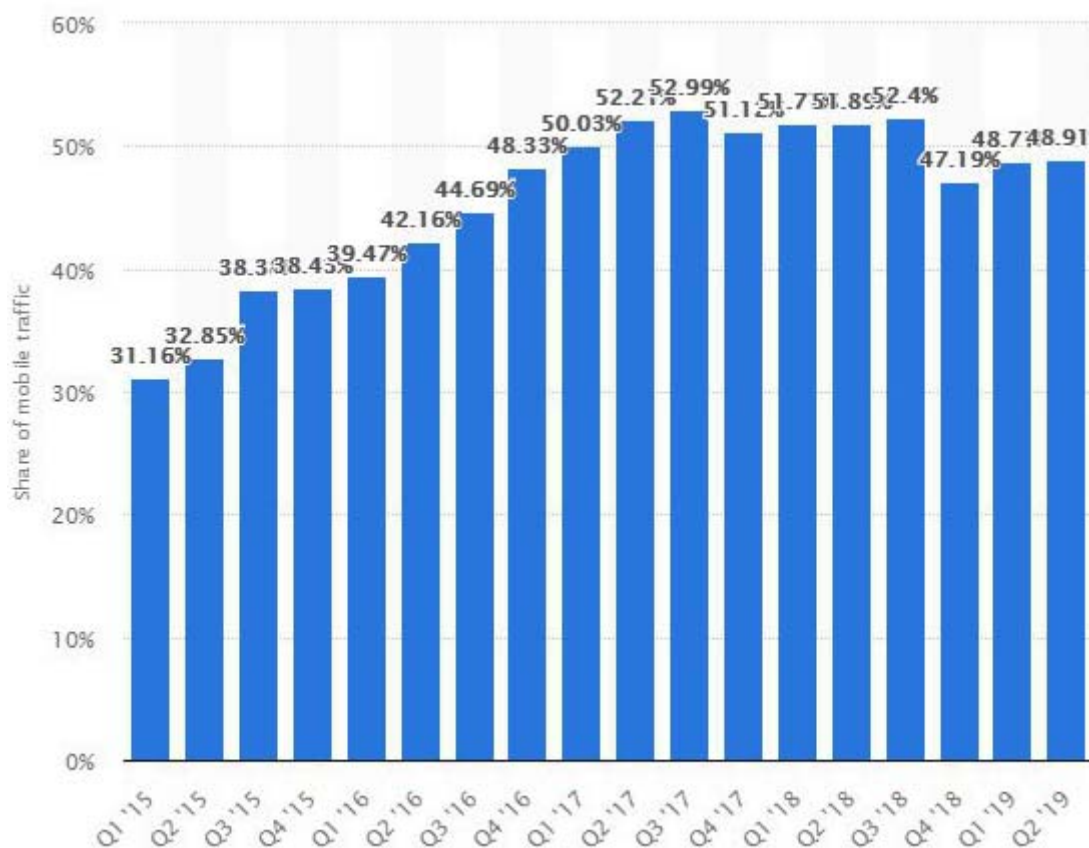


Figure 1. percentage of mobile site traffic



Figure 2. percentage of Suggested Crowdsourcing site traffic

In addition to the cost, another issue that was resolved in this plan was more use of mobile network in crowdsourcing process. In this study, by encouraging people in competition and planning to use people's free time, we were able to achieve 76% of mobile and tablet usage. People spend their leisure time with mobiles, the more we can orientate towards this group of society, the more we can utilize the capacity of mobile networks in crowdsourcing. Mobile network provides less user fatigue due to its ease of use and allows more crowdsourcing. According to Figure 1, the world averages mobile site traffic is up to 50%. This amount was 75% in our proposed method. (Figure 2) This indicates the maximum use of network capacity.

## 5. CONCLUSIONS

With the creation of text mining and data mining, a new window was opened to solve human beings' problems. Processes that were not previously understandable were investigated and the results were extracted from massive data. Feelings and opinions were assessed and the right move was made with the right roadmap. Scientists later found that many machine-based processes in data mining and sentiment analysis required human observers. An observer that improves the accuracy of the machine's performance. Sometimes this observer helped in the beginning of the activity and sometimes in results evaluation.

In this article, a plan is implemented to use the human interest in mining to use the mining process not as an abstract digital currency but as an actual good. All extracted goods are delivered to the extractors, who are the crowdsourcing workers, within one day of the month at Amir Kabir University of Technology. This process is repeated every month and therefore is not part of the ending processes. Creating a passion based on the mining process increased the use of this process in leisure time, as a result increased the use of mobile networking for crowdsourcing up to 76%.

## REFERENCES

- [1] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden, "Scaling up crowd-sourcing to very large datasets: a case for active learning," *Proceedings of the VLDB Endowment*, vol. 8, pp. 125-136, 2014.
- [2] Li, G., Wang, J., Zheng, Y., Fan, J., & Franklin, M. J. (2018). *Crowdsourcing Background*. In *Crowdsourced Data Management* (pp. 11-20). Springer, Singapore.
- [3] Correia, A., Schneider, D., Fonseca, B., & Paredes, H. (2018, September). *Crowdsourcing and massively collaborative science: a systematic literature review and mapping study*. In *International Conference on Collaboration and Technology* (pp. 133-154). Springer, Cham.
- [4] Steven C Brooks, Graydon Simmons, Heather Worthington, Bentley J Bobrow, and Laurie J Morrison, "The PulsePoint Respond mobile device application to crowdsource basic life support for patients with out-of-hospital cardiac arrest: Challenges for optimal implementation," *Resuscitation*, vol. 98, pp. 20-26, 2016.
- [5] Correia, A., Schneider, D., Fonseca, B., & Paredes, H. (2018, September). *Crowdsourcing and massively collaborative science: a systematic literature review and mapping study*. In *International Conference on Collaboration and Technology* (pp. 133-154). Springer, Cham.
- [6] Yashar Moshfeghi, Alvaro Francisco Huertas Rosero, and Joemon M Jose, "A Game-Theory Approach for Effective Crowdsourcing-Based Relevance Assessment," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, p. 55, 2016.
- [7] Miah, M. S. U., Masuduzzaman, M., Sarkar, W., Islam, H. M., Porag, F., & Hossain, S. (2018). *Intelligent Tour Planning System using Crowd Sourced Data*. *International Journal of Education and Management Engineering (IJEME)*, 8(1), 22-29.
- [8] Abhinav, K., Dubey, A., Jain, S., Bhatia, G. K., McCartin, B., & Bhardwaj, N. (2018, May). *Crowdassistant: A virtual buddy for crowd worker*. In *Proceedings of the 5th International Workshop on Crowd Sourcing in Software Engineering* (pp. 17-20). ACM.

- [9] Erik Cambria and Giuseppe Melfi, "Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis," in FLAIRS Conference, 2015, pp. 276-281.
- [10] Mingqing Hu and Bing Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.
- [11] Bing Liu and Lei Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data, ed: Springer, 2012, pp. 415-463.
- [12] K Veselovska, "Sentence-level polarity detection in a computer corpus," in Proceedings of the 20th Annual Conference of Doctoral Students. WDS, 2011, pp. 167-170.
- [13] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, vol. 28, pp. 15-21, 2013.
- [14] E Fersini, E Messina, and FA Pozzi, "Expressive signals in social media languages to improve polarity detection," Information Processing & Management, vol. 52, pp. 20-35, 2016.
- [15] Mehri Haghighi, "Data mining and machine learning: an Overview of Classifiers," Ciencia & Natura, vol. 37, pp. 76-86, 2015.
- [16] Cees GM Snoek and Marcel Worring, "Concept-based video retrieval," Foundations and Trends in Information Retrieval, vol. 2, pp. 215-322, 2008.
- [17] Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg, "Sentiment analysis using common-sense and context information," Computational intelligence and neuroscience, vol. 2015, p. 30, 2015.
- [18] US Patent and US Dep't of Commerce Trademark Office, "Manual of Patent Examining Procedure," ed, 2001.
- [19] Jacopo Staiano and Marco Guerini, "DepecheMood: a Lexicon for emotion analysis from crowd-annotated news," arXiv preprint arXiv:1405.1605, 2014.
- [20] Martin Jaggi, Fatih Uzdilli, and Mark Cieliebak, "Swiss-Chocolate: Sentiment Detection using Sparse SVMs and Part-Of-Speech n-Grams," in SemEval 2014-Proceedings of the 8th International Workshop on Semantic Evaluation, 2014, pp. 601-604.
- [21] Christopher H Lin and Daniel S Weld, "Re-active Learning: Active Learning with Relabeling," in Thirtieth AAAI Conference on Artificial Intelligence, 2016.

### AUTHOR PROFILE



Mohammad Jafaraabad received the B.S. and M.S. degrees in Computer Engineering from Islamic Azad University in 2007 and 2010, respectively. During 2012-2019, he is student of PhD program in Qom university. He is currently working on a Text Mining tool for analysing PhD Thesis.