Classification of Sentiment Analysis on Tweets Based on Techniques from Machine Learning

Anjali Singh[1], Mr.Manish Kumar Soni[2]

[1]M.Tech, Computer Science Engineering, Bansal Institute of Engineering & Technology, Lucknow, India
[2] Assistant Professor, Department of Computer Science Engineering, Bansal Institute of Engineering & Technology, Lucknow, India

---***---

Abstract - *This trend is expected to continue growing in popularity. When it comes to making specific choices, the public and the stakeholders might benefit from hearing the thoughts of the people. The process of retrieving information via the use of search engines, online blogs, micro-blogs, Twitter, and social networks is referred to as opinion mining. The user-generated material on Twitter provides a wealth of opportunities to collect the perspectives of people. It is challenging to physically delineate the information due to the enormous volume of tweets, which are in the form of unstructured text. Therefore, to mine and condense the tweets from the corpora, one has to be knowledgeable in computational techniques and have an understanding of phrases that convey emotions. The extraction of emotion from unstructured text may be accomplished by the use of a wide variety of computer models, methodologies, and algorithms. The vast majority of them are based on machine-learning strategies, with the Bag-of-Words (BoW) representation serving as the primary foundation. In this investigation, we used a lexicon-based technique to automatically identify sentiment for tweets that were gathered from the public domain of Twitter.*

Key Words: Bag-of-Words (BoW), Lexicon, Machine Learning Algorithms, Laplace Smoothing, Part-of-Speech (POS)

1. INTRODUCTION

In addition to this, an increasing number of people are turning to the internet to make it possible for people living in other countries to have access to their point of view. According to the findings of two surveys that were conducted on more than 2000 adults in the United States, each eighty percent of Internet users (or sixty percent of Americans) have conducted research on a product online at least once, and twenty percent (fifteen percent of Americans) prefer it on a specific day. These statistics were derived from the findings of two surveys that were carried out in the United States. The polls were conducted in the United States by two different organizations that are independent of one another. The United States of America served as the location for all of the polling and interviewing that took place. On the other hand, researchers working in the sector found the fact that around eighty thousand new blogs and two million new articles are generated every single day. Because of the widespread use of the internet and the continually shifting behaviors of customers, many of the conventional techniques of monitoring are becoming more useless. As a direct result of this, there is a desire for cutting-edge technology that is associated with product photographs for quality control. Therefore, in addition to individuals, a separate audience group for systems that can automatically analyze consumer sentiment, such as is described in no small part in online forums, are businesses that are eager to comprehend how their products and services are being perceived by customers in the market.

1.1. Principle of Sentiment Analysis

The source of the opinion, the target of the view, and the evaluative remarks or comments made by the opinion holder are the primary components that make up a SA problem. These components are not listed in any particular sequence. According to Liu [6,] the definition of a SA problem is as follows: "Given a set of evaluative text documents D that contain opinions (or sentiments) about an object, Sentiment Analysis aims to extract attributes and components of the object that have been commented on in each document d € D and to figure out whether the comments are positive, negative, or neutral."

1.2. Tracking down the Origins of Opinions

The person or medium that is responsible for the transmission of a feeling is considered to be the feeling's "carrier." The genesis of a feeling is the person or medium that is responsible for its transmission. The genesis of a feeling is of the utmost importance when it comes to validating and categorizing that feeling since it gives context to the emotion in the issue. This is because the origin of a position has a significant impact, both positive and negative, on the credibility and dependability of that viewpoint. This is because the origin of a perspective has a major impact on the quality of a feeling, which is expressed by that viewpoint. For instance, the weight of an expert's perspective may be compared to the weight of the opinion of the average person, and an opinion is considered to be credible when it is received from a respectable source.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

2. COLLECTION AND PREPROCESSING OF DATA

The rapid development of communication and information technology has resulted in the emergence of an entirely new set of challenges, which has greatly increased the level of difficulty associated with the process of information dissemination. One cannot resist emphasizing the relevance of social networks such as Facebook, Twitter, and MySpace while discussing the regulations that regulate the transmission of information and the collection of business intelligence. Examples of such networks are Facebook, Twitter, and MySpace. Tweets from Twitter were used to ensure that the behavior that was the subject of the study could be maintained. Twitter is unique among social networking sites in that each message may be no longer than 140 characters. This character limit sets Twitter apart from other social networking services. Despite this limitation, the information that is included in tweets is of immense value since it is feasible to extract a significant amount of data from such a restricted amount of space. Additionally, the images, videos, and transcripts of the presentations may all be seen together in one location, making it feasible to comprehend the whole narrative in a single sitting. The ease of access to the data is yet another crucial factor to take into account while making this decision. In the past, we were only able to put our model through its paces using a training set consisting of hundreds of objects. However, now that we have access to the Twitter API, we can gather millions of tweets to utilize as data for training our model. In the past, we were only able to put our model through its paces using a training set that had hundreds of different objects.

2.1. The Twitter API

When using the Twitter platform, users have access to two different kinds of application programming interfaces (APIs). These APIs are referred to as REST and Streaming respectively. In addition, the REST API is comprised of two more APIs, namely the REST API and the Search API (whose difference is due to their history of upgradation). The Streaming API, in contrast, to REST APIs, keeps an active connection to the server at all times and transfers data in a manner that is quite close to being in real-time. This differentiates it from REST APIs in a significant way. On the other hand, connections to REST APIs will only be permitted if they are active for a certain period and adhere to specified rate limitations (one can download a restricted amount of data per day). Users can access the data that Twitter has to offer at any time of the day or night thanks to the REST APIs.

2.2. From Twitter using Third Party API (Twython)

It is essential to collect only tweets that are specifically about that product or movie given that the purpose of the thesis is to determine the sentiment (positive, negative, or neutral) of tweets about a specific product or a movie, and it is the purpose of the thesis to determine whether tweets about the product or movie are positive, negative, or neutral. This is because the objective of the thesis is to analyze the sentiments expressed in tweets that are related to the product or movie in question. This is a result of the fact that the purpose of the thesis is to analyze the feelings that are expressed in tweets that are relevant to the product or movie that is under discussion. Having said that, putting this idea into effect won't be an easy task by any stretch of the imagination. It would seem that no method can be employed to use to collect all of the tweets that have been made about a certain subject that has been discussed.

Service	Term	Username	Name	Update	Link	Location	Followers	Friends	T
twitter	#joy	mariellabella	mariella	#pasta #meatballs #food #lunch #happy #mom #joy #yummy #handmac	http://twitter.	Trinidad	105	445	
twitter	#joy	KariJoys	Kari Joys MS	RT @DiRiseborough: #Letgo of the weight & find your #JOY on the	http://twitter.	Spokane, WA	23425	24309	
twitter	#joy	AllGodsThings	Omnia Dei	Alleluia! "@RadiateLA: #Easter is all about #joy. Radiate it! #RadiateLA	http://twitter.	Caribbean	266	338	
twitter	#joy	Rastamon86	fiestas EDM madrid	#Zoologicoclub #capital #joy #marcoAldani #madriz #salir #wekeend #d	http://twitter.	Madrid	1714	1366	
twitter	#joy	debsie301261	Deborah Myerscough	RT @ramblingsloa: " #Joy attracts #Joy" - Rhonda Byrne @thesecret @S	http://twitter.		525	581	
twitter	#joy	upfortomorrow	Tamara	RT @headquarters: I'm gonna root root root for BOTH teams! #Opening	http://twitter.		18	82	
twitter	#joy	wtjohnson01	Whitney Johnson	The little spaces we create for peace. #peaceful #spring #littlethings #l	http://twitter.		292	620	
twitter	#joy	ashrafxd	ashraf	RT @pixodentist: #FriFotos Blossom the #joy of Spring. http://t.co/7pA	http://twitter.	God`s own country / ma	314	1046	
twitter	#joy	LessysWorld	Lessy	And you should too! ?? #Joy #TrueJoyComesThroughChrist ? https://t.	http://twitter.	Brooklyn, NY	717	650	
twitter	#joy	eraivy	eraivy	RT @JohnSilkauskas: Start #singing start humming, start #worshiping ri	http://twitter.	kualalumpur	936	810	
twitter	#joy	NanaJenLaw	Nátalie Tavares	RT @Newsin01887: @JenLawUS @olv #Joy is filming RIGHT NOW at the	http://twitter.	RJ, Brasil	471	566	
twitter	#joy	patlejarde	Patricia Lejarde	luv u girlz @erikalaylo @ivymancilla @AsiRaihana @_coleenqb @jshxa	http://twitter.	PH	743	285	
twitter	#joy	TheJoyTrain	#JoyTrain	RT @DiRiseborough: #Letgo of the weight & find your #JOY on the	http://twitter.	Global	2269	2414	
twitter	#joy	lwrenci	amendoa	RT @Newsin01887: @JenLawUS @olv #Joy is filming RIGHT NOW at the	http://twitter.		1944	776	
twitter	#joy	bydefault904	Drizzy	@adamcarolla @Jokoy you don't understand Jo. We have a serious bro	http://twitter.	NE Florida	109	318	
twitter	#joy	kkax_	ลพคลmine	RT @parkkim9699: [CAP] 150404 #JOY ?http://t.co/pyallIp09N ©Studio_	http://twitter.	s k r 2 1 。	6039	131	
twitter	#joy	mathewdennis	Dennis Mathew	RT@Samsamum Focus for today Making peace with my complexity - #k	http://twitter.	Canada eh!	26284	26089	
twitter	#joy	TheJoyTrain	#JoyTrain	RT @KariJoys: @adservio Transform your #Anxiety to #Joy!#Spokane #J	http://twitter.	Global	2269	2414	
twitter	#joy	TheJoyTrain	#JoyTrain	RT @KariJoys: Real, Inspiring #Stories"People You May Know"Found #J	http://twitter.	Global	2269	2414	
twitter	#joy	xolovebeam	ปิ๊มสล็อต	RT @babyvelvet_th: [HQ] 150328 #JOY at Fansign Event In Busan ©mar	http://twitter.		625	1546	
twitter	#joy	TheJoyTrain	#JoyTrain	RT @KariJoys: Welcome to the #JoyTrain community! Just RT other twe	http://twitter.	Global	2269	2414	
twitter	#joy	TheJoyTrain	#JoyTrain	RT @KariJoys: #NewBook> "Who Stole Your #Joy?" Inspiring #hope!	http://twitter.	Global	2269	2414	
twitter	#joy	xolovebeam	ปิ๊มสล็อต	RT @babyvelvet_th: [OFFICIAL] 150407 #JOY & #JIMIN at After Sch	http://twitter.		625	1546	

Figure 1: Twitter data snapshot using a third-party API

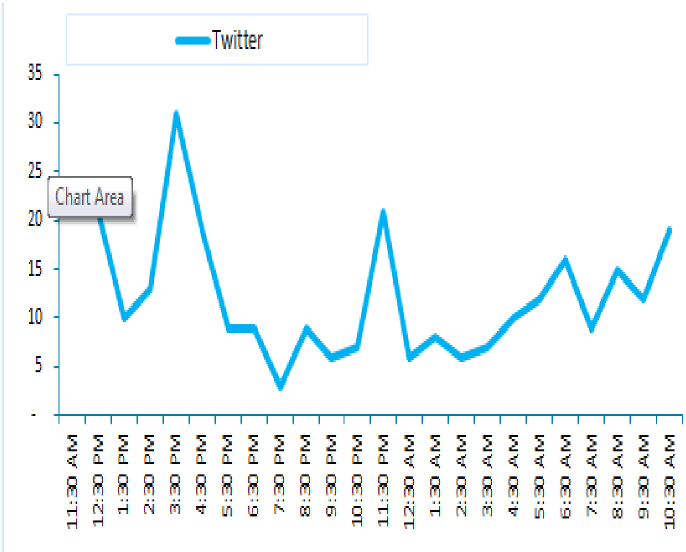

Figure 2: A snapshot of tweet volume relative to the period for "rage"

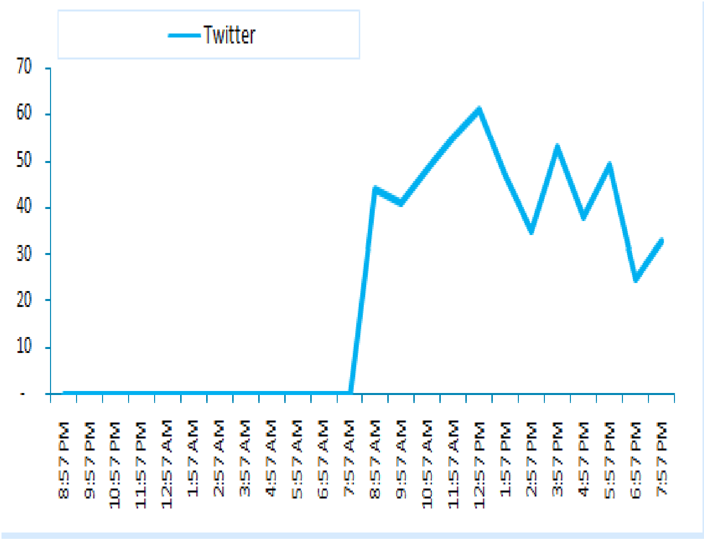

Figure 3: An overview of the number of tweets related to the hashtag "sad"

2.3. PREPROCESSING OF DATA

Preprocessing, which is an essential phase, is required to get rid of data that is either incomplete, noisy, or inconsistent. This is because preprocessing is a crucial step. This is necessary due to the need for preprocessing. Before using any of the data mining capabilities, it is necessary to do preprocessing in its entirety. This need must be satisfied. Before using any of the many methods that are available to us for doing sentiment analysis, we first carried out the preprocessing stage, which will be discussed in more depth in the following paragraphs (lexicon-based or machine learning).

2.3.1. Removal of the hash symbol

There is a certain label that goes by the name of a hashtag, and you can find it on social networking sites as well as other kinds of microblogging services. This form of label is used to categorize posts. Users are now in a position to identify messages that involve a certain topic matter or content in a far more expedient manner than was before feasible as a direct result of these labels. This advancement was previously impossible. For example, if you do a search using the hashtag #LOST (or #Lost or #lost as the search is not case-sensitive), we will get a list of tweets that are related to the hashtag #lost. These tweets will include references to #LOST.

3. LEXICON-BASED APPROACH IN SENTIMENT ANALYSIS

The use of terms that are thought to be indicative of either a positive or negative bias is taken into consideration to achieve the detection of subjectivity and the categorization of attitudes. This is done to accomplish both of these goals. The way of thinking that drives this technique is dependent on the idea that the meanings attributed to words may be understood in terms of the substance of views. This thesis serves as the foundation for this line of thought. The corpus of academic work that has been done in this field has documented the deployment of several beneficial techniques that are based on this strategy. These approaches have been successful in achieving their goals. For instance, Turney and colleagues [13] provide a method for the identification of subjectivity that is based on the use of a list of seed words that is determined by the proximity measure to other common phrases.

3.1. SentiWordNet

SentiWordNet is a lexical resource that was developed to aid with activities requiring opinion mining and sentiment analysis [14]. Its purpose was to help with these types of tasks. Its only objective was to assist in the aforementioned endeavors. An approach that is only semi-automatic is used to generate SentiWordNet's term-level opinion polarity, which is provided to users. The WordNet database, which contains English words and their relationships, is the source of the information that was utilized to generate this polarity.

3.2. WordNet

WordNet is a lexical database for the English language that was developed at Princeton University to realize the nature of semantic relations of terms in the English language. WordNet was developed as a lexical database for the English language because of the purpose of realizing these relations. Realizing the importance of these connections was the driving force for the creation of WordNet, a lexical database for the English language. The dawning understanding of the significance of these associations served as the impetus for the development of WordNet, a lexical database designed specifically for the English language.

3.3. Tagging of Speech Elements

The display of the information that can be accessed on SentiWordNet is in the hands of the POS, which is responsible for its proper organization. there are significant shifts that can take place in the amount of objectivity that a synset may reflect depending on the grammatical function that it plays. This can be seen as a direct result of the fact that there is a direct result of this. When classifying a source text, we need to extract the information on the parts of speech so that we may correctly apply the scores that we have acquired from the SentiWordNet database. This is required so that we can properly categorize the source text.

3.4. PROPOSED MODEL

In the last section of this chapter, the organization of the SentiWordNet database was dissected in great detail, and questions were raised about the challenges and constraints that are involved with the process of accumulating the required amount of opinion data. By making use of SentiWordNet, we were able to devise a model for the organization of features that have the potential to be used in the course of sentiment analysis. This is the location where you may find the model. The building of this model was accomplished while keeping all of these different issues in mind at the same time. However, the methodology for a list of features suggested in this section begins from the rule that the features obtained through SentiWordNet catch different aspects of tweet sentiment and are best suited to train any classifier algorithm. This rule serves as the foundation for the methodology for a list of features suggested in this section. The technique for a list of characteristics that are recommended in this part is built on top of this rule, which acts as the basis.

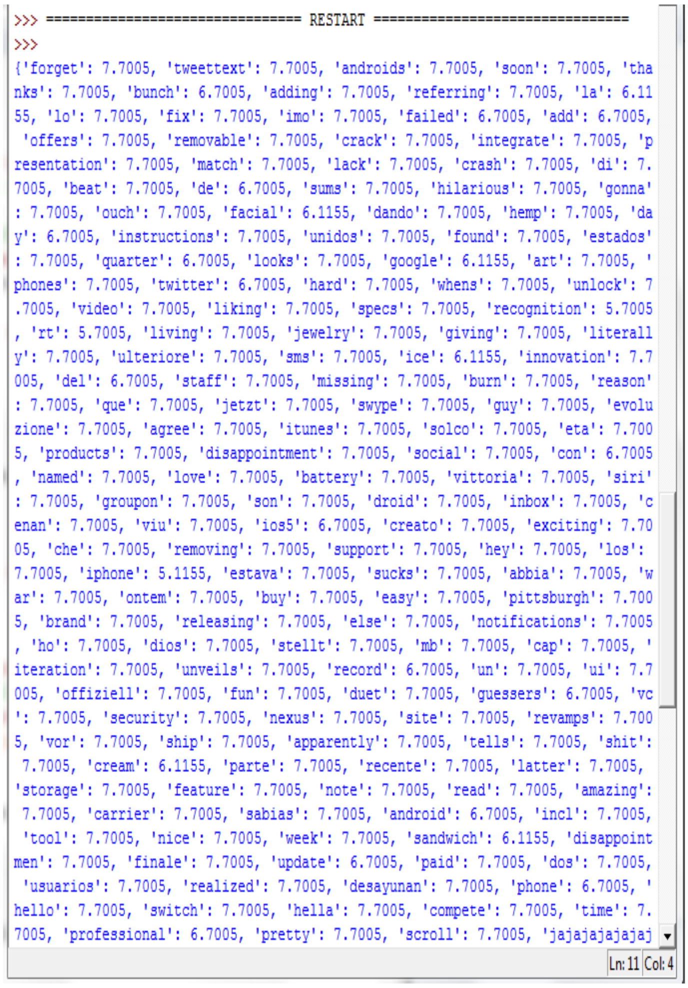

Figure 4: lexical snapshot produced from the tweet corpus

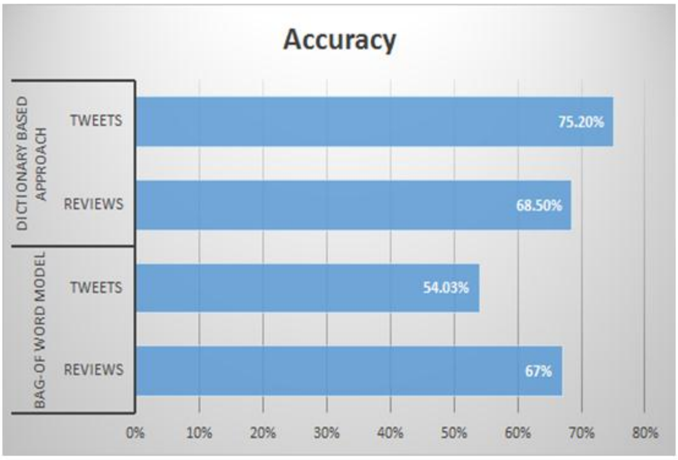

Figure 5: some quotes from a review of a film

4. INTERVENTIONAL SETUP

Figure 6 illustrates the component of our system that we have judged to be of the highest relevance, so go ahead and have a look at it. As we go ahead, we are going to discuss this issue in a more in-depth manner, focusing more on the details. To successfully carry out the implementation, we have taken use of our data collection, which is made up of 6000 tweets that have been sorted by hand into several categories. The compilation of tweets includes talks about businesses and services like Apple, Google, Microsoft, and Twitter, among others. Positive, negative, neutral, and irrelevant are the four classifications that may be applied to tweets. Tweets are considered to be unimportant if they are written in a language other than English or if there is no relationship between the tweet and the issue that is the focus of the conversation that is taking place right now. Throughout our investigation, we have placed the majority of our emphasis on three distinct categories: the positive, the negative, and the neutral.

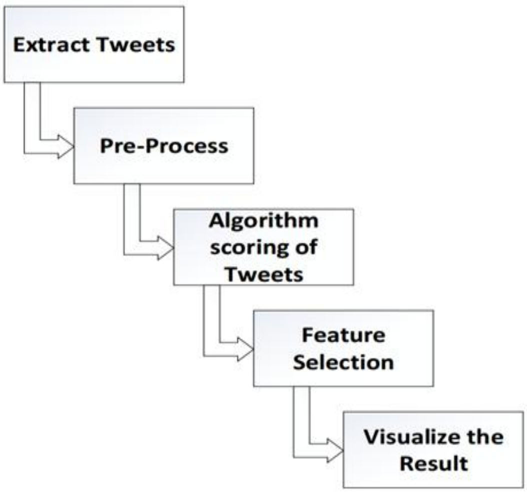

Figure 6: The experiment's intended block diagram

5. RESULT AND DISCUSSION

We investigate a wide range of different factors, each of which is known to have a significant impact on the findings of sentiment analysis. We made use of N-gram features such as unigrams (n = 1) and bigrams (n = 2), which are used often in a variety of text classifications including sentiment analysis. Specifically, we employed unigrams with n = 1 and bigrams with n = 2. We have focused specifically on unigrams and bigrams in our work. During our inquiry, we experimented with boolean qualities by using both unigrams and bigrams in our work. This allowed us to examine the relationship between the two. Each n-gram feature has a boolean value associated with it, which may be turned on or off according to the user's preferences. This value is only made to be true under the very precise condition that the required n-gram can be located in the tweet [12], which is a criterion that must be met.

Table 1: F1 rating for the MNB classifier

S.No	Class label	Precision(%)	Recall(%)	F1 score(%)
1	Positive	65.25%	20.51%	31.21%
2	Negative	77.41%	16.05%	27.20%
3	Neutral	80.48%	61.82%	69.93%

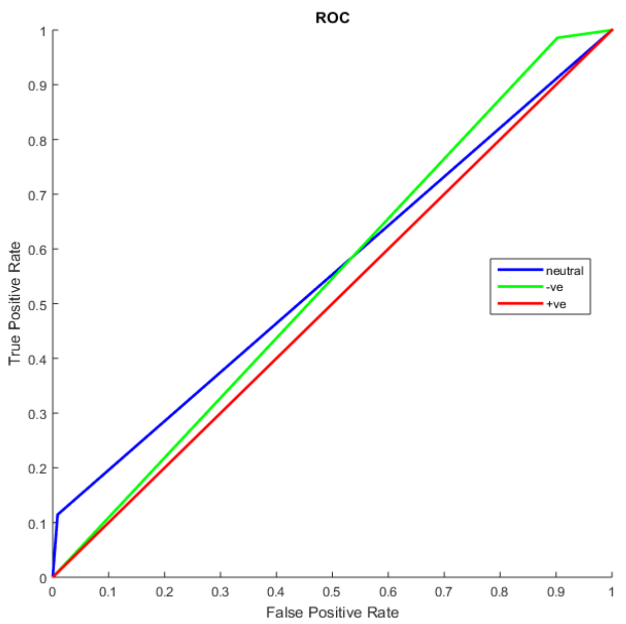

Figure 7: ROC curve for the MNB tweet classifier

The findings of our research indicate that the F1 score for the positive class and the negative class is much lower than the score for the neutral class. On the other hand, the score for the class that was seen as neutral was substantially higher. This is because the great majority of the tweets that were included in our data collection were manually categorized as terms that were either irrelevant or neutral. This is the reason why this is the case. The reasoning for why things are the way they are may be summed up as follows: As a direct result of this, to use any of the machine learning methods, we will need a training data set that is free from any kind of bias and does not call for any kind of human annotation. This will be necessary for us to be able to use any of the machine learning methods.

5.1. EMOTION DATASET

Hashtags are often used by people as a means of communicating the thoughts and feelings that are currently going on inside their minds. As a consequence of this, one may be able to deduce a sufficient number of ideas and feelings from the utterances that include the hashtag. To provide our machine learning algorithm with more data, we have upgraded it to take into consideration these hashtags. This will allow it to process more information.

~~~~
Confusion Matrix
             |                            t |
             |                            h |
             |                            a |
             |                            n |
             |                        s   k |
             |                        u   f |
             |                        r   u |
             |   a                    p   l |
             |   n   f       l        r   n |
             |   g   e   j   o   s    i   e |
             |   e   a   o   v   a    s   s |
             |   r   r   y   e   d    e   s |
-------------+------------------------------+
       anger |<195>  2   .   2   .    .   . |
        fear |   .<207>  3   .   3    1   . |
         joy |   .   1<189>  .   6    .   . |
        love |   1   .   4<211>  4    .   . |
         sad |   .   1   .   2<179>   .   . |
    surprise |   .   1   .   .   4<203>   . |
thankfulness |   .   6   9   5  10    .  <.>|
-------------+------------------------------+
(row = reference; col = test)

```
~~~~

Figure 8: Picture of the emotion dataset

6. CONCLUSION

This research is academic in the field of sentiment analysis; it focuses on the application of lexical resources and machine learning algorithms to the problem of classifying the emotional tone of tweets and text messages, two examples of unstructured data sources. Due to the abundance of subjective information accessible online, applications of Sentiment Analysis, which uses an automated approach to detect subjective content in a text, may be valuable in many sectors, including online advertising and market research. In the realm of knowledge management, this kind of opinion data is a crucial criterion since it is often the determining factor in major decisions. The reason we're doing this study is to learn more about the difficulties of Sentiment Analysis and the many approaches that have been developed to deal with them. Given the volume and variety of social media data, extracting its underlying sentiment might be challenging. Researching which elements are most useful for Sentiment Analysis, we picked tweets from the public stream and analyzed them. We have tried lexicon-based methods as well as machine-learning techniques for SA.

REFERENCE

[1] John A Horrigan. Online shopping. Pew Internet & American Life Project Report, 36, 2008.

[2] Bing Liu. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1):1–167, 2012.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1–135, 2008.

[4] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

[5] Bing Liu. Opinion mining and sentiment analysis. In Web Data Mining, pages 459–526. Springer, 2011.

[6] Bing Liu. Sentiment analysis and subjectivity. Handbook of natural language processing, 2:627–666, 2010.

[7] Andr´es Montoyo, Patricio Mart´ıNez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems, 53(4):675–679, 2012.

[8] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2):15–21, 2013.

[9] Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, and Ashraf Ullah. Mining opinion components from unstructured reviews: A review. Journal of King Saud University-Computer and Information Sciences, 26(3):258–275, 2014.

[10] Ronen Feldman, Moshe Fresko, Jacob Goldenberg, Oded Netzer, and Lyle Ungar. Extracting product comparisons from discussion boards. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pages 469–474. IEEE, 2007.

[11] Mohammad Sadegh, Roliana Ibrahim, and Zulaiha Ali Othman. Opinion mining and sentiment analysis: A survey. International Journal of Computers & Technology, 2(3):171–178, 2012.

[12] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.

13] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417–424. Association for Computational Linguistics, 2002.

[14] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC, volume 6, pages 417–422. Citeseer, 2006.

[15] Alaa Hamouda and Mohamed Rohaim. Reviews classification using sentiwordnet lexicon. In World Congress on Computer Science and Information Technology, 2011.

[16] Bruno Ohana. Opinion mining with the sentwordnet lexical resource. 2009.

[17] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. Computational linguistics, 19(2):313–330, 1993.

[18] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In Proceedings of the third conference on Applied natural language processing, pages 133–140. Association for Computational Linguistics, 1992.

[19] Shitanshu Verma and Pushpak Bhattacharyya. Incorporating semantic knowledge for sentiment analysis. Proceedings of ICON, 2009.

[20] Kamal Nigam, John La_erty, and Andrew McCallum. Using maximum entropy for text classification In IJCAI-99 workshop on machine learning for information filtering, volume 1, pages 61–67, 1999.

[21] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7):10760–10773, 2009.

[22] Daniel M Bikel and Je_rey Sorensen. If we want your opinion. In Semantic Computing, 2007. ICSC 2007. International Conference on, pages 493–500. IEEE, 2007.

[23] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 355–363. Association for Computational Linguistics, 2006.

[24] Matthias Hagen, Martin Potthast, Michel B¨uchner, and Benno Stein. Twitter sentiment detection via ensemble classification using averaged confidence scores. In Advances in Information Retrieval, pages 741–754. Springer, 2015.

[25] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. 2013.

[26] George Miller, Christiane Fellbaum, Randee Tengi, PWakefield, H Langone, and BR Haskell. WordNet. MIT Press Cambridge, 1998.

[27] Gebrekirstos Gebremeskel. Sentiment analysis of twitter posts about news. Sentiment Analysis. Feb, 2011.

[28] I Hemalatha, GP Saradhi Varma, and A Govardhan. Preprocessing the informal text for e_cient sentiment analysis. International Journal, 2012.