Deteksi Emosi Berbasis Teks Untuk Menganalisis Kuliah Daring Selama Masa Pandemi Menggunakan Algoritme K-*Nearest Neighbors*

Text Based Emotion Detection To Analyze Online Lecture During Pandemic Using K-Nearest Neighbors Algorithm

1st Mohammad Naufal Nabil Abdillah
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
naufalnabil@telkomuniversity.ac.id

2nd Casi Setianingsih
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
setiacasie@telkomuniversity.co.id

3rd Fussy Mentari Dirgantara
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
fussymentari@telkomuniversity.ac.id

Abstrak—Pada awal tahun 2020 terjadi sebuah peristiwa pandemi Covid-19, dimana instansi pendidikan memberlakukan kegiatannya secara online. Terdapat opini yang timbul di masyarakat terutama dari para mahasiswa yang mencurahkan emosinya di media sosial terutama Twitter. Penelitian ini bertujuan untuk mengetahui bagaimana emosi yang timbul dikalangan mahasiswa terkait dengan kuliah online. Pada Tugas Akhir ini digunakan algoritma K-Nearest Neighbor sebagai metode klasifikasi teks berbahasa Indonesia. Berdasarkan penelitian yang dilakukan oleh Shaver, terdapat lima kategori emosi dasar bahasa Indonesia yaitu marah, senang, sedih, takut, dan cinta. Pembagian data dibagi menjadi data tiga label emosi marah, senang, dan cinta, dan 4 label emosi marah, senang, takut, cinta. Data yang digunakan diambil dari scraping data twitter dan data Github. Pada Tugas Akhir ini, telah dilakukan pengujian menggunakan metode Confusion Matrix untuk mengetahui seberapa baik model yang telah dibuat pada sistem deteksi emosi berbasis teks. Hasil penelitian pada tugas akhir ini menunjukan bahwa sistem pendeteksi emosi berbasis teks dapat berjalan dengan baik dengan mendapatkan akurasi 78.91% pada data tiga label emosi pada partisi data 0.1, akurasi 69.74% pada data empat label emosi pada partisi data 0.2, dan akurasi 59.12% pada data lima label emosi pada partisi data 0.1.

Kata kunci : emosi, *k-nearest neighbor, text processing.*

Abstract—At the beginning of 2020 there was a COVID-19 pandemic, where educational institutions enforced their activities online. There are opinions from the public, especially from students who pour out their emotions on social media, especially Twitter. This study aims to find out how the emotions that arise among students are related to online lectures. In this final project uses the K-Nearest Neighbor algorithm as a method of classifying Indonesian texts. Based on according to a study conducted by Shaver, there are five basic categories of Indonesian emotions, namely anger, joy, sadness, fear, and love. The distribution of data is divided into three emotional labels of (anger, joy, and love), and four emotional labels of (angry, happy, fear, love).

The model that has been made is tested using the Confusion Matrix method to find out how well the model has been made to detect emotions

based on text. The results of this final project show that the text-based emotion detection system can run well by getting 78.91% accuracy on the three emotion label data on the 0.1 data partition, 69.74% accuracy on the four emotion label data on the 0.2 data partition, and 59.12% accuracy on the data partition. data of five emotional labels on 0.1 data partition.

Keywords: emotion, k-nearest neighbor, text processing.

I. PENDAHULUAN

Pada awal tahun 2020 telah terjadi wabah pandemi Covid-19, tepatnya pada bulan maret 2020 ditemukan kasus pertama Covid-19 di Indonesia. Kementerian Kesehatan telah merilis aturan turunan untuk merinci Peraturan Pemerintah (PP) Nomor 21 Tahun 2020 tentang Pembatasan Sosial Berskala Besar. Cakupan PSBB meliputi peliburan sekolah dan tempat kerja, fasilitas umum, kecuali supermarket, minimarket, pasar, toko, tempat penjualan obat-obatan dan peralatan medis, serta kebutuhan pokok [1].

Dengan diberlakukannya hal tersebut maka siswa maupun mahasiswa dituntut unutk melakukan kegiatan belajar mengajar dengan menggunakan sistem pembelajaran daring atau E-learning, yang mana muncul berbagai respon dari masyarakat terutama bagi siswa maupun mahasiswa yang tercurahkan pada media social seperti Twitter, Facebook dan lainnya.

Respon yang diberikan beranekaragam mulai dari yang merasa diuntungkan maupun yang merasa terbebani dari kebijakan tersebut, dari hal tersebut akan dilakukan sebuah penelitian dimana saya dan rekan saya akan mendeteksi emosi pada media teks dengan menggunakan pembelajaran mesin.

Penelitian ini dilakukan dengan tujuan untuk mengetahui emosi seseorang pada media teks, yang mana dapat menjadi masukan bagi instansi pendidikan dalam meningkatkan layanan dan mutu Pendidikan di Indonesia. Emosi menjadi salah satu poin penting bagi sebuah individu, emosi sudah tertanam dalam diri manusia sejak mereka lahir dan akan terus berkembang seiring dengan waktu.

Penelitian ini sudah banyak dilakukan tetapi menggunakan bahasa inggris dan masih sedikit yang menggunakan bahasa Indonesia. Dengan menggunakan data tweet berbahasa Indonesia yang membahas mengenai kegiatan kuliah online selama pandemi covid-19 di Indonesia. Untuk algoritma yang akan digunakan yaitu K-Nearest Neighbor dan hasilnya akan diimplementasikan kedalam website.

II. KAJIAN TEORI

A. Emosi

Emosi dapat diartikan sebagai reaksi yang dilakukan oleh tubuh dari situasi tertentu yang terjadi. Hal ini biasanya memiliki keterkaitan dengan aktivitas berpikir seseorang, dikarenakan hasil presepsi dari situasi yang sedang terjadi. Dalam sebuah jurnal yang dituliskan oleh Shaver dkk, analisis karakter data yang dikumpulkan di Indonesia mengungkapkan lima kategori emosi dasar diantaranya cinta, bahagia, marah, takut, sedih [2].

B. Web Scraping

Web scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, yang umumnya berupa halam-halaman web dalam bahasa markup seperti HTML, lalu menganalisis dokumen tersebut untuk mendapatkan data tertentu dari halaman tersebut untuk suatu kepentingan [3].

Web scraping hanya fokus pada cara untuk memperoleh data melalui pengambilan dan ekstrasi data dengan ukuran yang berfariasi. Salah satu software untuk webscraping yaitu Tweepy, Tweepy adalah paket Python *open-source* yang dapat mengakses API Twitter dengan Python.

C. Preprocessing Text

Text Preprocessing adalah sebuah tahap untuk melakukan seleksi data yang akan digunakan pada penelitian ini. Teks tidak bisa langsung diproses dengan menggunakan algoritma. Proses ini terdiri dari beberapa tahapan untuk membersihkan dokumen seperti *tokenizing*, *Slang Words, case folding*, dan *stemming* [4]. Adapun terdapat beberapa tahapan atau proses yang perlu dilakukan.

D. TF-IDF

Term Frequency Inverse Document Frequency (TFIDF) adalah sebuah metode pembobotan yang sering digunakan untuk ekstraksi data berupa teks. Tujuan dari metode pembobotan ini yaitu menemukan jumlah dari kata yang diketahui setelah

dikalikan dengan frekuensi kata tweet yang muncul, dengan cara integrasi antara term frequency dan inverse document frequency [5]. Berikut adalah rumus pembobotan dengan TFIDF.

$$tf(i) = \frac{freq_i\,(d_j)}{\sum_{i=1}^{k} freq_i\,(d_j)} \quad (1)$$

$$\text{IDFj} = 1 + \log\left(\frac{N+1}{DFi+1}\right) \quad (2)$$

$$\text{TFIDFij} = \text{TFi x IDFj} \quad (3)$$

E. K-Nearest Neighbors

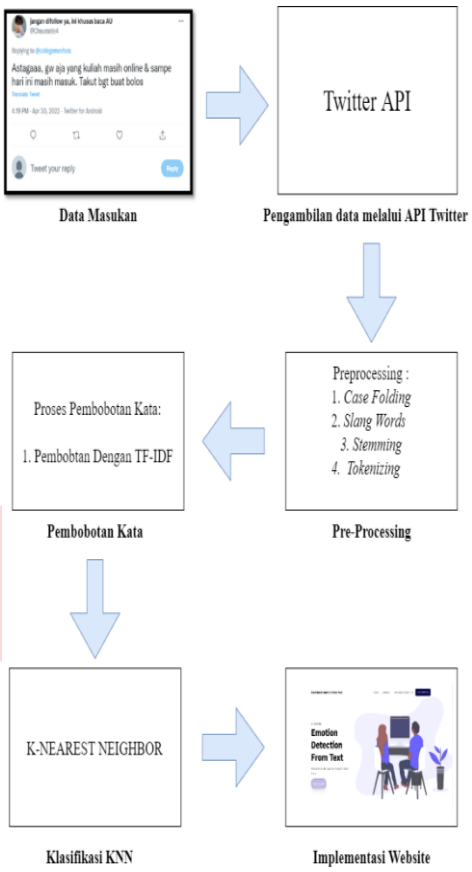
K-Nearest Neighbor adalah suatu metode yang digunakan dalam Data Maining. KNN menggunakan aturan jarak terdekat untuk melakukan pemilihan kelas, pemberian bobot berpengaruh terhadap jarak untuk memberikan pengaruh yang lebih kecil terhadap data yang terletak jauh dari data uji. Dalam melakukan prediksi data testing, data training akan mulai digunakan untuk mencari kemiripan data sesuai parameter K. Algoritma KNN mempunyai kelebihan yaitu mudah dimengerti dan diterapkan, proses pelatihan berlangsung cepat, bisa diimplementasikan pada kasus Multiclass [6]. Sedangkan kelemahan yang dimiliki KNN yaitu sensitif terhadap struktur data atau urutan data [7].

$$d = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2} \quad (4)$$

III. METODE

A. Gambaran Umum Sistem

Gambaran umum sistem deteksi emosi pada teks yang dibuat adalah sebuah aplikasi berbasis website yang akan memberikan sebuah keluaran jenis emosi pada kalimat berupa teks, terdapat lima emosi yang akan menjadi hasil dari sistem pendeteksi yaitu senang, sedih, marah, takut, dan cinta.

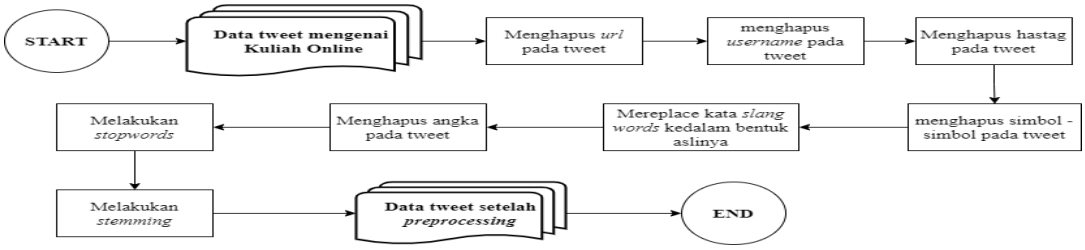


GAMBAR 1. Desain Sistem

Tweet pada Twitter akan diambil menggunakan API Tweepy dengan format data yang diambil adalah format CSV (COMMA Separated Values). Kemudian data akan diolah melalui beberapa tahapan yaitu pelabelan emosi secara manual oleh ahli bahasa yang dikategorikan berdasarkan pelabelan emosi oleh Shaver, dkk. Selanjutnya data akan melalui preprocessing dan pembobotan dengan TF-IDF sehingga data siap untuk diklasifikasikan dengan algoritma sebagai emosi yang akan digunakan user melalui user interface berbasis web.

B. *Text-Preprocessing*

Pada tahap pre-processing data akan diolah yang berfungsi untuk membersihkan data dan menghilangkan komponen yang tidak diperlukan yang dapat menyebabkan data sulit untuk diklasifikasi. Berikut merupakan alur tahapan dari proses pre-processing teks:



GAMBAR 2. Teks Preprocessing

1. *Case Folding dan Slang words*

Pada tahap ini dilakukan proses untuk mengubah semua huruf kapital menjadi huruf kecil, membersihkan teks dari teks – teks yang tidak diinginkan seperti angka, tanda baca, url, hastag karena hal tersebut tidak berpengaruh pada proses klasifikasi. Dan juga untuk slang words digunakan untuk mengubah kata tidak baku menjadi kata baku.

2. *Stemming*

Stemming merupakan proses transformasi kata yang terdapat dalam suatu dokumen menjadi kata-kata dasar dengan aturan tersentu, seperti pembuangan imbuhan pada kata dengan menggunakan bantuan tools Sastrawi.

3. *Tokenizing*

Tokenizing merupakan proses untuk memecah dokumen menjadi terms berdasarkan kata penyusunnya. Berikut contoh dari proses tokenizing.

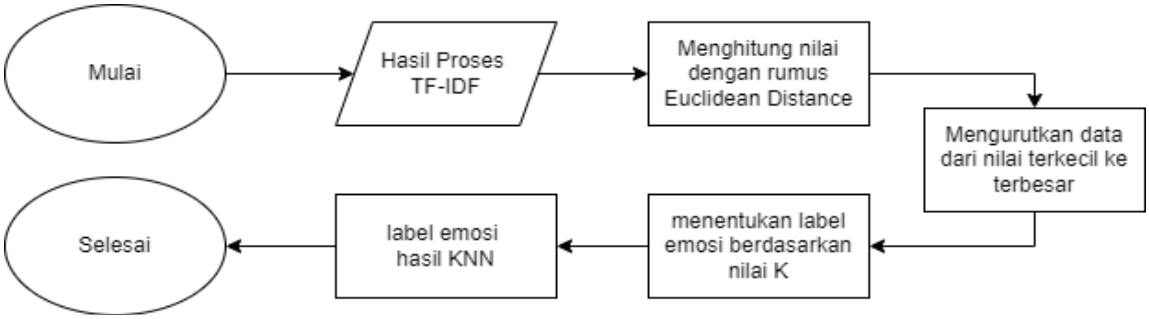
C. TF-IDF

TF-IDF merupakan proses perhitungan yang berfungsi untuk memberikan bobot nilai pada kata dengan cara melihat seberapa besar frekuensi kemunculan kata tersebut didalam sebuah dokumen. Berikut langkah dalam pembobotan dengan TF-IDF:

1. Memecah semua kalimat tweet menjadi kata yang berfungsi untuk mengetahui kata-kata yang muncul.
2. Menghitung jumlah kata yang terdapat pada setiap tweet.
3. Menghitung banyaknya kata dari suatu tweet menggunakan rumus 1.
4. Menghitung IDF dengan rumus 2 pada setiap kata yang muncul
5. Menghitung TFIDF dari setiap kata yang muncul pada sebuah tweet menggunakan rumus 3.

D. K-Nearest Neighbors

Pada pembuatan tugas akhir ini algoritma klasifikasi yang digunakan yaitu K-Nearest Neighbor. Pada metode ini dilakukan klasifikasi objek ke dalam salah satu kelas terdekat dari kelas yang telah ditetapkan dari label emosi yang telah dibuat secara manual yaitu cinta, marah, senang, sedih, takut.



GAMBAR 3. Proses Klasifikasi KNN

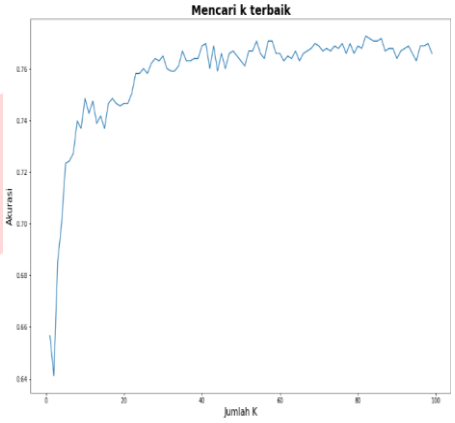
Berikut merupakan penjelasan berdasarkan urutan prosesnya. Untuk menentukan nilai K dimulai dari K=1. Nilai K akan ditambah sampai mendapatkan nilai kesalahan minimum atau nilai akurasi tertinggi. Data yang digunakan yaitu frekuensi kata dari setiap kata yang muncul pada tweet. Antara data training dan testing dihitung menggunakan jarak eucliden dengan rumus (4).

Data yang didapat dari hasil perhitungan nilai jarak terdekat akan diurutkan dari nilai yang terkecil, sehingga diperoleh data jarak terdekat berdasarkan nilai dari proses sebelumnya. Dan Pengklasifikasian kelas dilihat dari nilai jarak antara data training dan testing berdasarkan nilai K. frekuensi tiap kelas dan data training akan diklasifikasikan kedalam kelas dengan nilai mayoritas.

IV. HASIL DAN PEMBAHASAN

A. Pengujian Nilai K Terbaik

Pada pengujian nilai K terbaik dari algoritma k-NN menggunakan 3 dataset yang memiliki 3 label emosi, 4 label emosi dan 5 label emosi. setiap dataset dibagi menjadi 80% data latih dan 20% data uji.

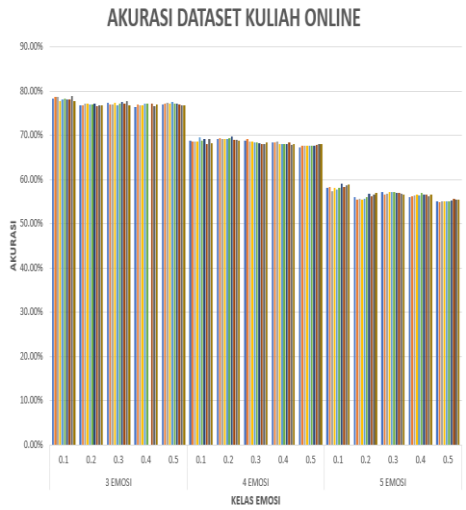


GAMBAR 4. Grafik nilai K =1-100

Tahapan pada pengujian ini yaitu menggunakan nilai K=1 sampai dengan K=100, kemudian hasil dari akurasi pada nilai K=1 sampai dengan K=100 dimunculkan dalam bentuk grafik. Maka dapat dilihat nilai K terbaik pada grafik tersebut sebagai acuan nilai terbaik.

B. Pengujian Partisi Data

Pengujian ini akan melakukan partisi data menjadi beberapa bagian. Pengujian akan dilakukan sebanyak lima kali, dengan tiap pengujian akan menguji partisi data yang berdeda, berikut merupakan data yang didapatkan dengan penampilan pada nilai k terbaik pada setiap pengujian partisi.



GAMBAR 5. Grafik hasil Seluruh Pengujian partisi

Pada gambar 5 dapat dilihat hasil akurasi partisi terbaik yaitu pada data 3 label emosi dengan partisi data uji 10% dengan akurasi 78.91%. dapat disimpulkan semakin sedikit emosi maka semakin tingggi akurasi yang didapat.

C. Pengujian *Confusion Matrix*

Pada Confusion Matrix terdapat nilai True Positive, False Positive, True Negative, dan False Negative. Nilai tersebut nantinya akan digunakan untuk melihat keberhasilan sistem berdasar besar akurasi, presisi, recall, dan F1 score yang didapatkan.

TABEL 1. Confusion Matrix

Dataset Kuliah Online 3 label K = 38 Partisi Data 90% Train 10% Test					
		Label Kelas Hasil Klasifikasi			Total
		Marah	Senang	Cinta	
	Marah	174	17	4	195
	Senang	42	121	30	193
	Cinta	7	9	113	129
Total		223	147	147	517

TABEL 2. Hasil *Confusion Matrix*

Presisi Per Kelas Emosi		Recall Per Kelas Emosi	
Emosi	Presisi	Emosi	Recall
Marah	0.78027	Marah	0.89231
Senang	0.82313	Senang	0.62694
Cinta	0.76871	Cinta	0.87597
Akurasi		78,91%	
Presisi		79%	
Recall		79.84%	
F1 Score		78.77%	

Pada tabel 1 menunjukan angka matrix hasil klasifikasi yang digunakan untuk menghitung akurasi, presisi, recall, dan F1 score. Pada tabel 2 menunjukan hasil yang diperoleh dari hasil perhitungan menggunkaan matrix.

D. Pengujian K-Fold Cross Validation

Tujuan pengujian ini untuk melihat kinerja dari algoritma K-Nearest Neighbor dengan menggunakan nilai K terbaik. Proses cross validation dilakukan dengan membagi data dengan cv = 10. Berikut pengujian yang dilakukan dengan k-Fold Cross Validation:

Dataset Kuliah Online													
Jumlah Label	K	CV										Mean	Akurasi
		1	2	3	4	5	6	7	8	9	10		
3	38	74%	79%	77%	75%	78%	75%	75%	78%	76%	78%	76%	78%
4	86	65%	69%	70%	65%	67%	65%	66%	65%	65%	68%	67%	69%

Dataset Kuliah Online													
Jumlah Label	K	CV										Mean	Akurasi
		1	2	3	4	5	6	7	8	9	10		
5	46	55%	59%	56%	54%	54%	56%	57%	55%	56%	52%	56%	59%

Pada table 4.7 dapat dilihat bahwa rata-rata akurasi yang didapat dengan menggunakan pengujian K-Fold Cross Validation dengan nilai CV = 10 tidak terlalu jauh hasilnya dengan akurasi yang diperoleh dari hasil klasifikasi menggunakann algoritma K-Nearest Neighbor. Hal ini dapat diartikan bahwa algoritma yang digunakan sudah berjalan dengan baik.

V. KESIMPULAN

A. Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan pada Tuags Akhir ini, maka dapat disimpulkan sebahai berikut:

1. Sistem deteksi emosi pada teks dalam bahasa Indonesia menggunakan algoritma K-Nearest Neighbors berbasis website telah berhasil dalam melakukan klasifikasi emosi terkait kuliah online.
2. Pada Pengujian yang telah dilakukan didapatkan hasil akurasi tertinggi pada dataset kuliah online dengan 3 label emosi dengan nilai akurasi = 78.91%, presisi = 79%, *recall* = 79.89% dan F1 *Score* = 78.77% dengan nilai K = 38 dan partisi data 90% data latih 10% data uji.
3. Hasil yang didapat pada pengujian *K-Fold* tidak berbeda jauh dengan hasil *confusion matrix*.

B. Saran

Adapun saran dari penulis untuk pengembangan penelitian di masa yang akan datang, yaitu:

1. Melakukan klasifikasi dengan algoritma deep learning untuk mendapatkan hasil preformansi sistem yang lebih baik.
2. Perlu dilakukan pencarian data tweet lebih dalam untuk kelas label emosi takut dan perlu dilakukan pembersihan dan perubahan kembali bahasa daerah yang terabawa saat proses pengambilan data ke bahasa Indonesia.

REFERENSI

[1] S. Achmad Syauqi, "JALAN PANJANG COVID19 (sebuah refleksi dikala wabah merajalela berdampak pada perekonomian)," *JKUBS*, vol. 1, pp. 6-8, 2020.

[2] P. R. S. a. U. Murdaya and R. C. Fraley, "Structure Of The Indonesian Emotion Lexicon," pp. 201-224, 2001.

[3] T. A. H. Dana Febri Setiawan, "APLIKASI WEB SCRAPING DESKRIPSI PRODUK," *Jurnal TEKNOINFO*, Vols. Vol. 14, No. 1, pp. 41-42, 2020.

[4] L. Hermawan and M. B. Ismiati, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *TRANSFORMATIKA*, pp. 188-199, 2020.

[5] A. C. F. A. P. A. A. A. S. H. D. F. Moh. Afif Rofiqi, "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *Journal of Computer Science and Applied Informatics*, Vols. Vol. 1, No. 2, pp. 58-64, 2019.

[6] A. G. Hetal Bhavsar, "Variations of Support Vector Machine classification Technique: A survey," *International Journal of Advanced Computer Research*, Vols. Volume-2 Number-4,

pp. 255-256, 2012.

[7] E. P. K. Arifin, "Classification of Emotions in Indonesian Texts Using K-NN Method," *International Journal of Information and Electronics Engineering*, Vols. Vol.2, No.6, 2012.

[8] P. Ekman, "Basic emotions. In T. Dalgleish & M. J. Power (Eds.)," *Handbook of Cognition and Emotion*, pp. 45-60, 1999.

[9] a. N. S. Mariko Shirai, "Is Sadness Only Onse Emotion? Psychological and Physiological Responses to Sadness Induced by Two Different Situations: "Loss of Someone" and "Failere to Achieve a Goal"," *Frontiers in Psychology*, vol. volume 8, pp. 1-10, 2017.

[10] J. Aronoff, "How We Recognize Angry," *Cross-Cultural Research*, vol. Vol. 40 No. 1, pp. 13-14, 2006.

[11] R. Williams, "Anger as a Basic Emotion and Its role in Personality Building and Pathological Growth: The Neuroscientific, Developmental and Clinical Perspectives," *Frontiers in Psychology*, vol. volume 8, pp. 1-9, 2017.

[12] K. A. L. a. L. F. Barrett, "Constructing Emotion The Experience of Fear as a Conceptual Act," *Association for Psychological Science*, vol. Volume 19—Number 9, pp. 898-903, 2008.

[13] H. J. M. A. S. W. PHILLIP R. SHAVER, "Is love a “basic” emotion?," *Personal Relationships*, pp. 81-96, 1996.